Semantic Modeling and Knowledge Representation in Multimedia Databases

Wasfi Al-Khatib, *Member*, *IEEE Computer Society*, Y. Francis Day, *Member*, *IEEE*,
Arif Ghafoor, *Fellow*, *IEEE*, P. Bruce Berra, *Fellow*, *IEEE*

Abstract—In this paper, we present the current state of the art in semantic data modeling of multimedia data. Semantic conceptualization can be performed at several levels of information granularity, leading to multilevel indexing and searching mechanisms. Various models at different levels of granularity are compared. At the finest level of granularity, multimedia data can be indexed based on image contents, such as identification of objects and faces. At a coarser level of granularity, indexing of multimedia data can be focused on events and episodes, which are higher level abstractions. In light of the above, we also examine modeling and indexing techniques of multimedia documents.

Index Terms—Multimedia data modeling, multimedia document modeling, semantic modeling, knowledge representation, content-based retrieval, image databases, video databases, video indexing, query formulation, query processing.

1 Introduction

Multimedia databases have been the subject of extensive research during the last ten years. Interest in this area is rapidly growing with the continuous evolution of telecommunication and computing technologies. A number of applications in telemedicine, digital libraries, distance learning, tourism, distributed CAD/CAM, GIS, etc. are expected to use general purpose multimedia database systems. With the rapid proliferation of the Web, these applications are rapidly emerging. Unlike traditional alphanumeric databases, multimedia databases require manipulation of complex objects consisting of text, images, graphics, audio, music, and full-motion video data.

The objective of this paper is to highlight issues in the design and development of such systems. The primary emphasis is on semantic modeling of multimedia information, indexing, and knowledge based representation of semantics associated with image, video, and complex multimedia documents.

Data abstraction is an essential component of the modeling process, and requires the use of knowledge-based representations and spatio-temporal semantics. Such representations can vary in terms of handling precision and fuzziness in query processing, ranging from very specific to very general. We discuss several approaches in this area in the following sections, and provide a critical assessment of the comparative applicability of these approaches under disparate real-world applications.

In addition to the management of individual media, it is necessary to develop models for composing complex multimedia objects and documents. Composition can be both in space and time. Such models require a tight linking with the underlying image, audio, text and video database management systems. The objective of spatio-temporal modeling is two fold. Firstly, it can lead to the design of efficient retrieval algorithms for various multimedia data types which comprise the documents. Secondly, it can provide the basis for developing indexing and searching of these documents. Integration of these models with higher level information abstractions such as hypermedia, or object-oriented models is required to allow users to search and browse this large repository of data. Throughout this paper, special consideration is given to integration issues as they relate to the unique features of multimedia documents.

This paper is organized as follows. In the next section, we discuss issues in data modeling and knowledge representation for image databases. Section 3 tackles these issues in relation to video databases. In Section 4, issues pertaining to multimedia document modeling and management are discussed. A summary of the paper is given in Section 5.

2 Semantic Modeling and Knowledge Representation in Image Databases

Research in image database systems has traditionally focused on the design of robust image processing and recognition techniques. The growing role of images in multimedia applications has spurred tremendous interest in management aspects of image databases. Challenges inherent in the development of image databases include image processing for feature extraction, identification of salient objects, development of efficient data models to allow content-based indexing and retrieval, query formulation, and fuzzy

- *W. Al-Khatib and A. Ghafoor are with the School of Electrical and Computer Engineering, Purdue University, 1285 Electrical Engineering Building, West Lafayette, IN 47907. E-mail: {wasfi, ghafoor}@ecn.purdue.edu.*
- *Y.F. Day is with the Multimedia Documentation Program, Siemens Corporate Research, 755 College Rd. E., Princeton, NJ 08540. E-mail: fday@scr.siemens.com.*
- *P.B. Berra is with the Department of Computer Science and Engineering, Wright State University, Dayton, OH 45435. E-mail: bberra@cs.wright.edu.*

Manuscript received 12 June 1997; revised 6 Nov. 1998.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 109083.

query processing. Most traditional approaches to image data management use multilevel abstraction mechanisms to support content-based retrieval. These levels are depicted in Fig. 1 and correspond to three key functionalities: feature extraction, object recognition, and domain-specific spatial reasoning and semantic modeling. This figure provides the focus of our discussion. We use it for elaborating key issues and approaches proposed in the literature related to the development of multilevel abstraction and indexing mechanisms. The important role played by knowledge-based representation in processing queries at different levels is also discussed. In Section 2.1, functionalities of the feature extraction layer are discussed. Section 2.2 describes the functionalities of the object recognition layer. In Section 2.3, we discuss the models and approaches used for semantic and knowledge representation.

2.1 Feature Extraction Layer

Low-level image processing involves finding portions of the raw data which match the user's requested pattern. During such processing, features in the user's requested pattern need to be identified and matched against features stored in the database. In other words, pattern matching needs to be employed to find portions of the image that are similar to a given pattern or to deformed versions of a given pattern. In this context, deformations of the pattern include scaling, shifting, rotation, and stretching of the given pattern.

The problem of pattern matching in image databases has been actively studied for over 20 years [19], [36]. A typical approach is to extract a set of features which concisely describes the given pattern, and then seek these features in the image. Extraction of feature vectors has been well studied in certain specialized settings such as face recognition [9] and character recognition [24]. However, only recently have the researchers started addressing the problem of automatic feature extraction for a multimedia database for a wide variety of objects [33]. Image features include colors, textures, shapes, edges, and the like. These features are mapped into a multidimensional feature space which can allow similarity based retrieval of images. Features in an image can be classified as global or local. Global features generally emphasize "coarse-grained" pattern matching techniques. The global feature extraction techniques transform the whole image into a "*functional representation.*" In this case, finer details within individual parts of the image are ignored. Color histograms, Fast Fourier Transform, Hough Transform, and Eigenvalues are the well-known functional techniques that fall into this category. Fig. 2 is an example of a typical histogram of an image. Such a histogram can be stored as an approximate signature for the image, and used in subsequent pattern matching.

Example queries involving global features include

- "*Find images which are predominantly green,*" or
- "*Retrieve an image with a large round orange textured object*."

Global features for image indexing carry out uniform processing over the whole image for the chosen feature. Because of this, global features are well-suited for processing the type of queries which deal with images as single entities during the pattern matching process. Such techniques are also useful for comparing images or video frames to identify changes in the global features in order to detect scene change, as discussed in Section 3.1. Major advantage of the *coarse-grained* approach is the low computational complexity of feature extraction and pattern matching algorithms. The disadvantage is that a high percentage of irrelevant images maybe retrieved as a result of query processing. Improvement in the precision and accuracy of the retrieval process is possible by incorporating textual annotation with images [30].

At another level of granularity, local features can be used to identify salient objects in an image and to extract more information about the finer details in the image. This approach is called "*fine-grained*" because images are segmented into a collection of smaller regions, with each region representing a potential object of interest. An object of interest may represent a simple semantic object, such as a "*round object*" or "*a region with uniform color.*" Segmented regions are processed to extract multiple features. Local

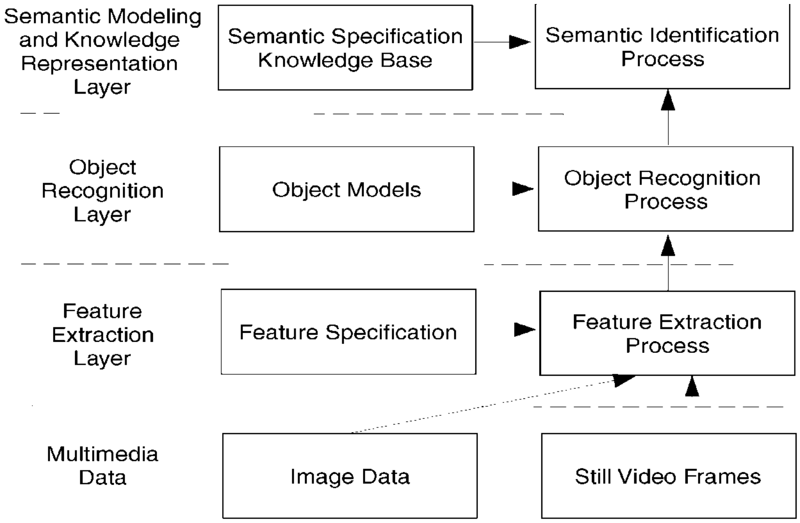

Semantic Modeling and Knowledge Representation Layer
Semantic Specification Knowledge Base
Semantic Identification Process
Object Recognition Layer
Object Models
Object Recognition Process
Feature Extraction Layer
Feature Specification
Feature Extraction Process
Multimedia Data
Image Data
Still Video Frames

Fig. 1. Processing and semantic modeling for image database.

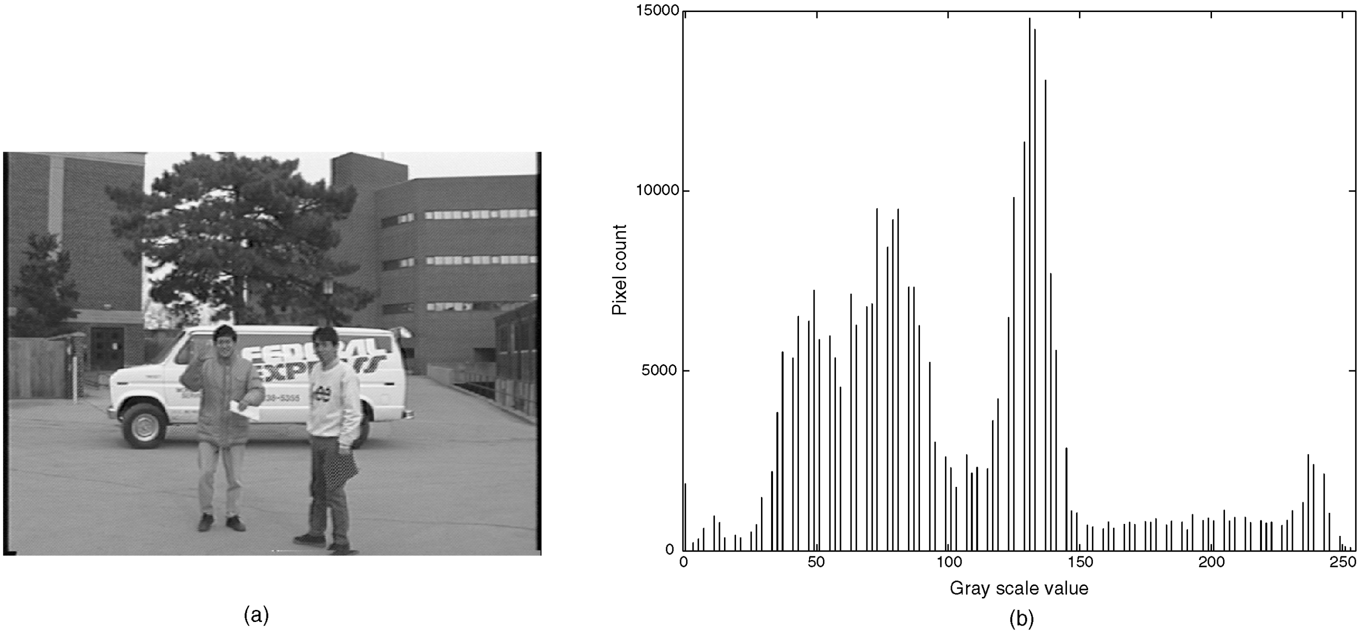


Fig. 2: (a) (640 × 480) gray scale image; (b) its color histogram.

features can be perceived as constituting a multidimensional search space. Features in the form of encoded vectors provide the basis for indexing and searching mechanisms of image databases. Typical features include grey scale values of pixels, colors, shapes, and texture. Various combination of features can be specified at the time of formulating database queries. Segmented regions are further analyzed by the object recognition layer to identify objects with higher level semantics (Section 2.2).

Incorporating knowledge about objects with local features can provide more robust and precise indexing and search mechanisms. Measures such as Minkowski distance [38], weighted distance [16], [37], color histogram intersection [41], and average distance [26], [34] can be used to evaluate the robustness of the mechanisms. The performance of a similarity-based search depends on the degree of imprecision and fuzziness introduced by the types of features used, and the computational characteristics of the search algorithm. Faster algorithms require dropping more image features and hence provide less precise results. The trade-off is between the time complexity and robustness of the algorithms. Choice of features, their extraction mechanisms and the search process are domain specific. For example, multimedia applications targeted for X-ray imaging, GIS, etc., require spatial features such as shapes and dimensions. On the other hand, color features are more suitable in applications involving MMR imaging, paintings, etc.

Various systems have been prototyped that use a feature extraction layer similar to the one shown in Fig. 1. For example, in the Query By Image Content (QBIC) system [14], color, shape, and texture features are used for image retrieval. A fully automatic image segmentation method is used to identify objects in images with few foreground objects on separable background. The QBIC system allows querying of the database by sketching features and providing color information about the desired objects.

Chabot [30] is another system which uses a combination of color and textual annotation attributes for image retrieval. Chabot uses the notion of "*concept query*" where a concept, like sunset, is recognized by analyzing images using color features. It uses a frame-based knowledge representation of image contents which is precomputed and stored as attributes in a relational data model. For improving the performance of the system, it uses textual annotation of images by keywords that are manually entered.

Hsu, Chu, and Taira use quantitative methods for edge detection to identify shape features in a radiological database, KMeD [17]. KMeD employs a three layer architecture, where the lowest layer, known as the representation layer, uses shapes and contours to represent features. This layer employs a semiautomatic feature extraction mechanism based on a combination of low level image processing techniques and visual analysis of the image manually. From a functionality point of view, this layer reduces to the feature extraction layer of Fig. 1.

2.2 Object Recognition Layer

Features are analyzed by this layer to recognize objects and faces in an image database. The process involves matching features with the object models stored in a knowledge base. In general, an object model is a template describing a specific object. During the matching process, each template is inspected to find the "closest" match. Object identification with an exact match is generally computationally expensive and the quality of matching depends on the details and the degree of precision provided by the object template. Occlusion of objects and the existence of spurious features in the image can further diminish the success of matching strategies.

Two types of template matching have been proposed in the literature: fixed template matching and deformable template matching [35]. Approaches based on fixed templates are useful when object shapes do not change with respect to the viewing angle of the camera. Image subtraction and correlation have been used in fixed template matching techniques. In image subtraction techniques, the difference in intensity levels between the image and the template is used in object recognition. The template position is determined from minimizing the distance function between the template and various positions in the image. Although image subtraction techniques require less computation time than correlation techniques, they perform well in restricted environments where imaging conditions, such as image intensity between the template and images containing this template are the same. An example application where subtraction technique is appropriate is an X-ray image database, since target images have a fixed viewing angle and image intensities do not change substantially.

Matching by correlation utilizes the position of the normalized cross-correlation peak between a template and an image to locate the best match. This technique is generally immune to noise and illumination effects in the images, but suffers from high computational complexity caused by summations over the entire template. Point correlation can reduce the computational complexity to a small set of carefully chosen points for the summations [22].

Deformable template matching approaches are more suitable for cases where objects in the database may vary due to rigid and nonrigid deformations. In this approach, a template is represented as a bitmap describing the characteristic contour/edges of an object shape. A probabilistic transformation on the prototype contour is applied to deform the template to fit salient edges in the input image [19]. An objective function with transformation parameters which alter the shape of the template is formulated reflecting the cost of such transformations. The objective function is minimized by iteratively updating the transformations parameters to best match the object. Applications of deformable template matching techniques include handwritten character recognition and motion detection of objects in video frames.

A certain level of fuzziness and imprecision in object recognition is inevitable and needs to be incorporated in the similarity measure in order to increase the success rate of queries and not to exclude good candidates. For this reason, manual examination of output images is generally unavoidable. For example, medical objects belonging only to patients in a small age group, are identified automatically in KMeD [17]. In addition, such objects have high contrast with respect to their background and have relatively simple shapes, large sizes, and little or no overlap with other objects thus resulting in relatively accurate matching results. The domain knowledge of KMeD maintains descriptions about the shapes of objects with simple structures such as tumor, brain, bones, etc. For large age groups, however, there are few objects with simple structures [17]. KMeD resorts to a human-assisted object recognition process in such cases.

Identification of human faces is another important application of image databases. However, due to more inherent *structuredness* in human faces, models and features used for face recognition are different than those used for object recognition. Face recognition involves three steps: face detection to locate a face inside an image; feature extraction where various parts of a face are detected; and face recognition where the person is identified by consulting a database containing *facial models*. Several face detection and recognition systems for multimedia environments have been proposed [28]. Most of these systems use information about various prominent parts of a face such as eyes, nose and mouth for face recognition. Other techniques decompose face images into a set of characteristic features called eigenfaces [28]. These techniques capture variations in a collection of face images and use them to encode and compare individual features. Other techniques employ a profile-posed approach, where detailed structure of the face not seen in frontal images is stored. In particular, the size and orientation of the nose is delineated. Face recognition process in this class of techniques is based on profiles and it concentrates on locating points of interest, called *fiducial points*, and determining the relationships among these fiducial points.

Extraction of features and object recognition are important phases in developing large scale general purpose image database management systems. Significant results have been reported in the literature for the last two decades, with successful implementation of several prototypes. However, lack of precise models for object representation and the high complexity of image processing algorithms make the development of fully automatic image management and content-based retrieval systems a challenging task.

2.3 Spatial Modeling and Knowledge Representation Layer

The major function of this layer is to maintain the domain knowledge for representing spatial semantics associated with image databases. Queries at this level are generally descriptive in nature, and focus mostly on semantics and concepts present in image databases. For most of the applications, semantics at this level are based on “spatial events” [13] describing the relative locations of multiple objects. Such semantics are used for high-level indexing and content-based retrieval of images. An example involving such semantics is a range query which involves spatial concepts such as *close by, in the vicinity, larger than*, etc. The most common applications employing spatial semantics and content-based retrieval based on range queries are map databases and geographic information systems (GIS). This type of systems is extensively used in urban planning and resource management scenarios. In clinical radiology applications, relative sizes and positions of objects are critical for medical diagnosis and treatment. Some example queries in this application include:

- “*Retrieve all images that contain a large tumor in the brain,*” or
- “*Find an image where the main artery is 40 percent blocked.*”

Such techniques are aimed at inferring new information pertaining to the evolutionary nature of the data.

The general approach for modeling spatial semantics for such applications is based on identifying spatial relationships among objects once they are recognized and marked by the lower layer using bounding boxes or volumes. Spatial relationships can be coded using various knowledge-based techniques. Several techniques have been proposed to formally represent spatial knowledge at this layer. These include: semantic networks, mathematical logic, constraints, inclusion hierarchies, and frames. Brief descriptions of these approaches are given below.

2.3.1 Semantic Networks

Semantic networks are used extensively in artificial intelligence applications. They were first introduced to represent the meanings of English sentences in terms of words and relationships between them [42]. This is a graph-based approach to represent spatial concepts and relationships. Semantic networks are graphs of nodes representing concepts that are linked together by arcs representing relationships between these concepts. Efficiency in semantic networks is gained by representing each concept or object once and using pointers for cross references rather than naming an object explicitly every time it is involved in a relation. Further extensions of semantic networks can allow efficient search strategies to locate the desired information. One such system that uses hierarchical semantic networks is the KMeD system [17] discussed in the previous sections. In the second layer of this system, known as the semantic layer, objects and their relationships are identified and abstracted using an Entity-Relationship model. The third layer of this system uses a knowledge base abstraction to represent higher level semantics [11]. This abstraction is called type abstraction hierarchy (TAH) that is organized using semantic networks. TAHs conceptualize the objects and their semantics and incorporate the domain expert knowledge in order to improve search efficiency in radiological databases.

2.3.2 Constraints-Based Methodology

In this methodology, the domain knowledge is represented using a set of constraints in conjunction with formal expressions such as predicate calculus or graphs. Knowledge is represented by predicates that are augmented with procedural information. A *constraint* is a relationship between two or more objects that needs to be satisfied. An example of this approach is the PICTION system [40]. Its architecture consists of a natural language processing (NLP) module, an image understanding module (IU), and a control module. This architecture is functionally similar to that given in Fig. 1. The system uses a combination of text annotation and image processing techniques for face recognition and identifying relative positions of people in images. In this approach a set of constraints is derived by the NLP module from the picture captions. These constraints, also termed as *visual semantics* are used with the faces recognized in the picture by the IU module to identify the spatial relationships among people. The control module maintains the constraints generated by the NLP module and acts as a knowledge-base for the IU module to perform face recognition functions. Fig. 3a shows an example image with caption: “In front of the electrical engineering building, Francis Day is standing to the right of Jaehyung Yang.” The result of employing image processing techniques that locate possible candidates for faces is shown in Fig. 3b. Applying natural language processing to the caption generates the constraint graph shown in Fig. 3c. Fig. 3d shows the result of applying the constraints to the candidate faces. Such a system is suitable for those image database applications where sufficient descriptive information of images is available. An important application of this methodology is in the management of captioned news image / video databases.

2.3.3 Mathematical Logic

Mathematical logic provides powerful techniques to formulate knowledge representation. In multimedia databases, techniques based on such representations can be used to represent high level spatial semantics [3], [8], [13]. For example, the approach in [8] uses projections of salient objects in a coordinated system. These projections are expressed in the form of 2D strings to form a partial ordering of object projections in 2D. In other words, these expressions characterize the spatial relationships among objects. For query processing, 2D subsequence matching is performed to allow similarity-based retrieval. Several spatial propositions used in GIS applications can be used to identify spatial contents in an image, including: near, far, inside, above, below, aligned and next [3]. Imprecision in these prepositions is handled through a fuzzy function which allows range-based specification of spatial relations.

Day et al. describe a framework that uses the notion of binary spatial relations [13]. A set of 13 relations is used to specify such propositions. These relations, shown in Fig. 4, have been originally proposed for temporal reasoning [5]. They consist of 13 relations, and can be represented by seven operators because six of them have inverses. For example, **after** is the inverse relation of **before**. For inverse relations, given any two intervals, it is possible to represent their relation by using the noninverse relations only by exchanging the interval labels. The equality relation has no inverse.

2.3.4 Inclusion Hierarchies

Another category of knowledge representation is *inclusion hierarchies* which group together semantically related objects. This approach is oriented in flavor and uses concept classes and attributes to represent domain knowledge [45]. These concepts may represent image features, high-level semantics, semantic operators and conditions. The operators and conditions are used to formulate descriptive queries involving features and range of values for objects. Being a hierarchical formalism, this approach renders itself nicely for object oriented modeling.

2.3.5 Frames

Frames are used to represent knowledge relevant to particular objects, situations, or concepts. A frame usually consists of a name and a list of attribute-value pairs. A frame can be associated with a class of objects or with a class of concepts. A frame-based approach to represent knowledge through media abstractions is proposed in [7]. Frame abstractions allow encapsulation of file names, features, and relevant attributes of image objects.

2.4 Summary and Challenges

Table 1 summarizes the characteristics of several prototyped image database systems. Their key features are highlighted. One observation from this table is that the underlying design philosophy of these systems is driven by the application

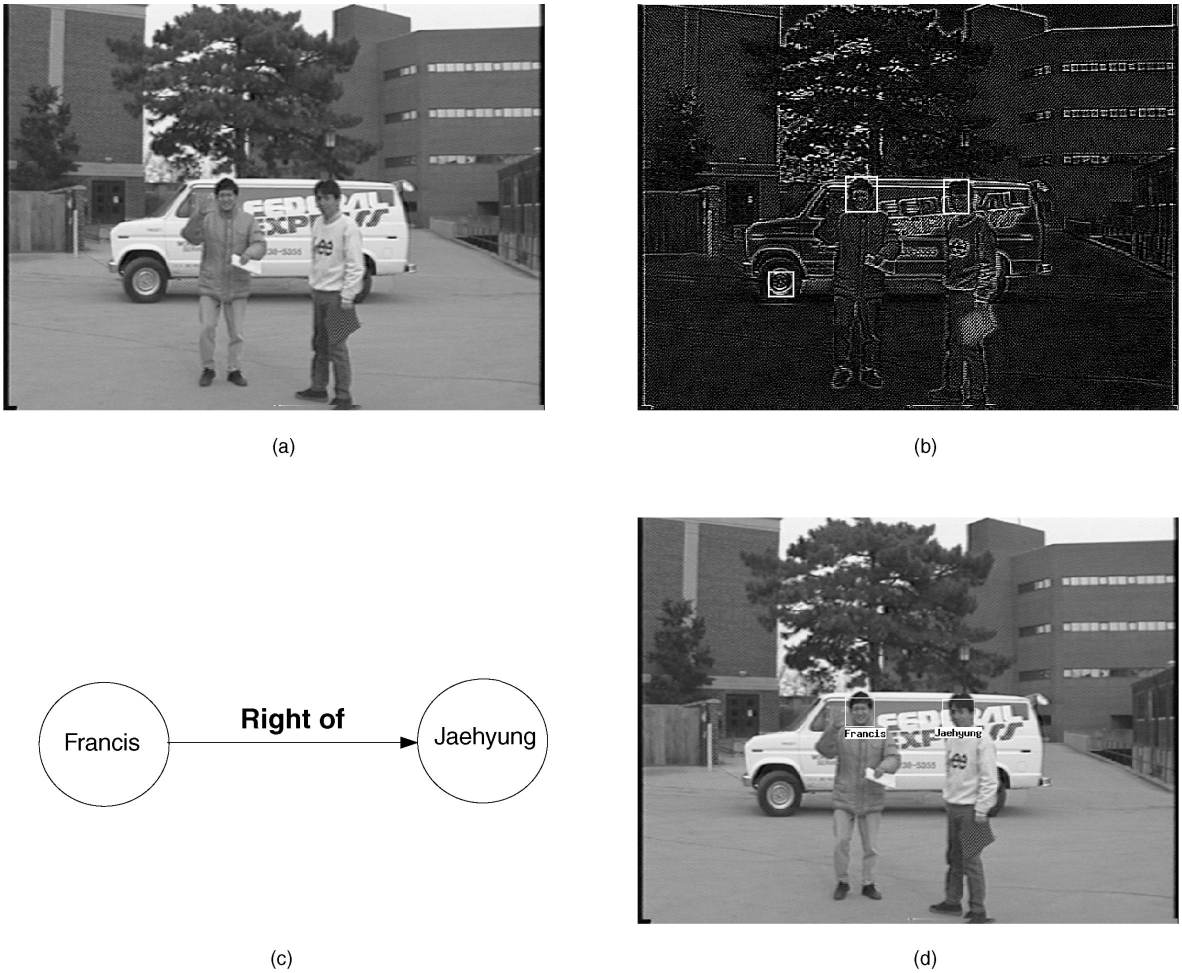


Fig. 3: (a) Original image with the following accompanying caption: “In front of the electrical engineering building, Francis Day is standing to the right of Jaehyung Yang;” (b) a sample edge image where image processing techniques are used to locate faces; (c) constraint graph generated from natural language processing of the caption; (d) final pictorial output with faces labeled by their respective names.

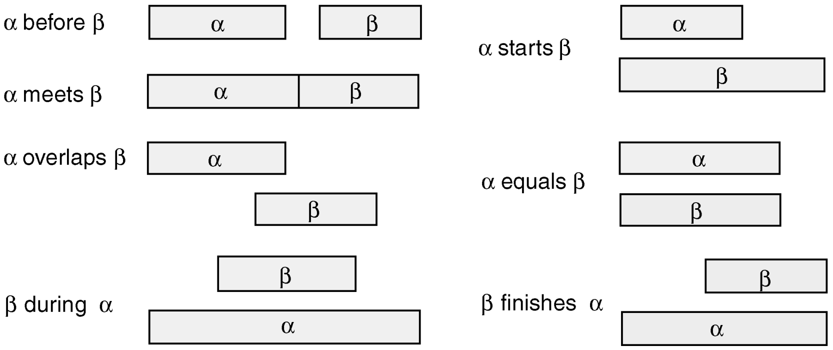


Fig. 4. Binary relations.

domain. Development of a general-purpose, automatic image database system capable of supporting arbitrary domains is a challenging task due to the limitations of existing image processing knowledge representation models.

We conclude this section by listing several challenges that need further investigation. There are three major areas of research identified with image databases regarding semantic modeling and knowledge representation: development of efficient knowledge representation schemes, development of efficient image retrieval algorithms and modeling of higher level semantics of images. Extracting knowledge from images and representing such knowledge is a challenging problem and requires further research. Extracting domain-specific features that facilitate the modeling of higher level semantics is crucial in improving the performance of object and pattern recognition techniques. The current limitations of image processing technologies and lack of precise query formulation mechanisms

TABLE 1
HIGHLIGHTS OF SEVERAL IMAGE DATABASE SYSTEMS

System	**Layer**					
	Feature Extraction		Object Recognition		Spatial Semantics	
	Process	Features	Process	Type of Knowledge Base	Process	Knowledge Base Support
QBIC [14]	Automatic	Color, Shape	Hybrid	–	–	–
Chabot [30]	Automatic	Color	Keywords	Frame Based	–	–
KMeD [17]	Hybrid	Shape	Hybrid	Attribute List of shape descriptors	Hybrid	Semantic Nets
PICTION [40]	Automatic	Facial Shape	Automatic	Constraints	Automatic	Constraints
Yoshitaka et al. [45]	Automatic	Shape	Manual	–	Automatic	Inclusion Hierarchies

introduce high degrees of imprecision in the retrieval process for image databases. This problem can be acute for large-scale databases. Alternatively, such limitations can be partially overcome by resorting to human intervention and manual annotation.

3 SEMANTIC MODELING AND KNOWLEDGE REPRESENTATION IN VIDEO DATABASES

The key characteristic of video data that makes it different from a-isochronous data such as text, image, and maps is its temporal dimension. In addition to spatial semantics, video events generally have high degree of temporal contents. An example of a video query involving spatio-temporal semantics is:

- *"Find video clips with a touchdown event."*

Important considerations in video data modeling are specification of such semantics and development of indexing mechanisms. Another critical issue is to cope with semantic heterogeneity that may arise due to differences in the interpretations of information in a video clip by different sets of users. Semantic heterogeneity has proven to be a difficult problem in conventional databases, with little or no consensus on the way to tackle it in practice. In the context of video databases, the problem becomes more difficult and intractable.

Generally, semantics and events in video data can be expressed in terms of interplay among physical objects in space and time. For data modeling purposes, spatio-temporal relations among objects can be represented in a suitable indexing structure, which can then be used for query processing.

In this section, we describe issues in event-based semantic modeling and knowledge representation of video data. We consider two criteria for classifying existing approaches of modeling video data. Based on these two criteria, we identify five major classes of approaches that are being employed in modeling video data, as shown in Fig. 5. According to the first criterion, semantic modeling is classified based on the level of abstraction of the identified events. The level of abstraction is considered *low* if the supported semantics are low-level semantics which are more relevant to the machine than users. An example of low-level semantics is scene changes. The level of abstraction increases as the degree of information contents and knowledge extracted from video data increases. For example, "*scoring a field goal*" in a sports video data carries a high level of information. However, any camera breaks introduced in the data may provide only a different viewing angle of the football field. In the second approach, the emphasis is on the method of preprocessing of video data which is the basis on which semantics are extracted. The preprocessing is of *coarse granularity* if it involves processing of video frames as a whole, whereas it is of *fine granularity* if processing involves detection and identification of objects within a video frame. Models classified on the right hand side of Fig. 5 are considered of *fine granularity* since they focus on processing

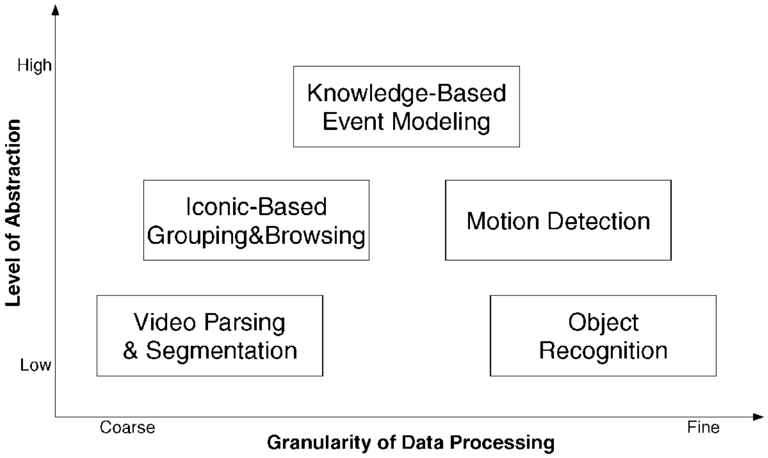

High
Low
Level of Abstraction
Knowledge-Based Event Modeling
Iconic-Based Grouping&Browsing
Motion Detection
Video Parsing & Segmentation
Object Recognition
Coarse
Fine
Granularity of Data Processing

Fig. 5. Semantic modeling of video data.

video data at the object level, whereas the focus of models on the left-hand side is on processing video data at frame level using global features. In the rest of this section, we discuss basis for classification of these approaches and the functionalities of each technique.

3.1 Video Parsing and Segmentation

Video parsing employs image processing techniques to extract important global features from individual video frames. Any significant change in the feature value in a sequence of frames is used to mark a change in the scene. The process allows a high level segmentation of video data into several shots. Several scene change detection methods have been proposed in the literature. These include: pixel-level comparison, likelihood ratio, color histogram, χ^2 histogram, and discrete cosine transform (DCT) based approaches for compressed video data.

In the pixel-level comparison approach, gray-scale values of pixels at corresponding locations in two distinct frames, either consecutive or a fixed distance apart, are subtracted and the absolute value is used as a measure of dissimilarity between the pixel values [29], [46]. If this value exceeds a certain threshold, then the pixel grey scale is assumed to have changed. The percentage of pixels that have changed is the measure of dissimilarity between the frames.

The pixel-level comparison approach is sensitive to several factors, including noise introduced by the digitization process, object movement and camera effects. A modification that can be used to limit the effect of such problems is to subdivide the frame into regions and select only certain regions for processing. In this approach, known as the likelihood ratio approach [20], [46], the frames are divided into blocks. The blocks of two consecutive frames are compared based on some statistical characteristics of their intensity values, such as the mean value of intensity. This approach is more robust than the pixel-level comparison approach in the presence of noise and object movement.

In the color histogram approach [29], [46], a frame is analyzed by dividing the color space into discrete colors called bins and counting the number of pixels that fall into each bin. A separate histogram is made for *R*, *G*, and *B* components of colors present in a frame. A variation of histogram technique uses a normalization approach [29] that results in large differences being made larger and small differences being made smaller. In compressed domain, discrete cosine transform (DCT) approach has recently been proposed that carry the advantage of being less costly in computation [27], [47]. Table 2 provides a summary of these approaches in terms of their relative advantages and limitations.

3.2 Iconic-Based Grouping and Browsing

Parsed video segments can be grouped together based on some similarity measure of image features possessed by one or more frames representing a scene, also known as *representative frames*. This can be used to build iconic based browsing environments. In this case, a *representative frame* of each scene is displayed to the user in order to provide the information about the objects and possible events present in that scene. An example along these lines is described in [44]. A directed graph is used to portray an overall “visual” summary of a video clip consisting of different scenes that may appear more than once in the video sequence. Nodes represent representative frames and edges denote the temporal relationships between them, giving an overview of the order in which these events have occurred. In another approach, a similarity pyramid is utilized to give a hierarchical clustering of all representative frames present in the video database [10]. Organization of this pyramid is based on extracted features and user’s feedback. This scheme can be scaled up to manage large numbers of video sequences and can provide browsing environments for developing digital video libraries.

3.3 Object Recognition

The main function of the object recognition layer is to identify key objects and faces and perform motion analysis to track their relative movements. The granularity of data processing at this level is fine-grained since such processing involves recognition of objects in individual video frames. For this purpose, each video frame is examined either manually or analyzed using image processing techniques for automatic recognition of objects and faces, as discussed in Section 2. Furthermore, motion information has been used to identify key objects in video sequences [4]. In this

TABLE 2
ADVANTAGES AND LIMITATIONS OF SEVERAL VIDEO PARSING TECHNIQUES

Technique	Advantages	Limitations
Pixel level comparison	able to detect sudden appearance and disappearance of objects	sensitive to noise, change in illumination and camera movement
Likelihood ratio comparison	able to detect sudden appearance and disappearance of objects, less prone to noise	sensitive to change in illumination and camera movement
Color histogram	robust in presence of camera or object motion	heavily dependent on pixel intensities
χ^2 color histogram	more robust than color histogram	–
Discrete Cosine Transform (DCT) based parsing	computationally efficient, suitable for compressed video	–
Inner product with DCT coefficients	computationally efficient, suitable for compressed video, less prone to noise	–

approach, motion and temporal information can be combined with classical image processing techniques to produce robust results allowing detection of objects without requiring any a priori assumptions.

3.4 Motion Detection

A major challenge in video data modeling is capturing of motion information about salient objects and persons. We need to extract information at higher levels of abstraction of identified objects from a sequence of frames related to the motion of these objects. Incorporating of knowledge of background information in tracking object motion results in a coarser granularity of data processing than that of *object recognition* as shown in Fig. 5. Several approaches have been proposed in the literature for tracking motion of objects. In this section, we elaborate on two techniques. The first technique uses a modified version of known compression algorithms such as MPEG, to identify and track motion of salient objects. In essence, this semantic-based compression approach combines both image processing and image compression techniques. In [15], for example, a motion tracking algorithm uses both forward and backward motion vectors of macroblocks used by MPEG, encoding algorithm to generate trajectories for objects. These trajectories are subsequently used by the higher layer for motion-based semantic modeling.

The second approach for motion tracking uses a directed graph model to capture both spatial and temporal attributes of objects and persons. The model, known as Video Semantic Directed Graph (VSDG), is used to maintain temporal information of objects once they are identified by image processing techniques [13]. This is achieved by specifying the changes in the parameters of a 3D projection associated with the bounding volume of objects in a given sequence of frames. At the finest level of granularity, these changes can be recorded for each frame. Although such a fine-grained motion specification may be desirable for frame-based indexing of video data, it may not be required in most of the applications requiring temporal modeling. In addition, the overhead associated with such detailed specification may be formidable. Alternatively a coarse-grained event specification can be generated by analyzing selected frames for motion tracking at some fixed distance apart. Such skip distance may depend upon the complexity of the event. There is an obvious trade-off between the amount of storage needed for event specification and the detailed information maintained by the model. Both these approaches and several others can be used to describe higher level events using knowledge-bases, as discussed in the following section.

3.5 Knowledge-Based Event Modeling

Based on the information available from the low layers of Fig. 5, higher level events (events that are meaningful to users) can be specified by the user to construct different views of the video data. Modeling at this level combines coarse information and fine details of video frames and is placed in the middle of the *granularity of data processing* classification axis. There has been a growing interest in using knowledge-based techniques to model high level semantics and events in video data [6], [10], [12], [13], [15], [31], [39], [43].

In order to develop high level semantics based on video parsing and segmentation, scenes are clustered together, automatically or manually, based on some desired semantics. There are several ways to build such abstractions. For example, clustering can be based on key objects and other features within each scene which are identified using image processing techniques, or textual information from video caption, in case it is available. Such an approach is proposed [10] where a set of features of video frames along with the motion vectors are used. The approach classifies the data into pseudo-semantic classes corresponding to head and shoulders, indoor versus outdoor, high action, and man-made versus natural [10]. In another clustering approach, domain specific semantics in form of sketches or *reference frames* are used to identify video segments that are closely related to these frames. Smoliar and Zhang [39] use this approach by exploiting the well-structured domain of news broadcasting to build an a priori model consisting of *reference frames* as a knowledge base that assists in semantic-based classification and clustering of video segments related to news broadcast. Alternatively, the scenes of the segmented video can be examined manually in order to append appropriate textual description. Such description can then be used to develop semantic-based clustering by examining events present in different scenes.

Temporal modeling has been extensively used to capture knowledge and semantics in temporal databases. Several approaches recently proposed in the literature develop knowledge-based formalisms for event specification of video data [6], [12], [13], [15], [31], [43]. Semantic operators including logic, set, and spatio-temporal operators, are extensively used. Logical operators include the conventional Boolean connectives such as *not*, *and*, *or*, *if-then*, *only-if*, and *equivalent-to*. Set operators like *union*, *intersection*, and *difference* are mostly used for event specification as well as for video composition and editing [31], [43]. The crux of this formalism is a set of spatio-temporal operators, based on temporal relations for event specification and modeling [6], [13], as shown in Fig. 4. In essence, the approaches proposed in the literature use different combinations of these operators. In the following section, we provide an overview of temporal intervals and discuss their role in developing high level video modeling.

3.5.1 Overview of Temporal Modeling

Temporal modeling of is used to construct complex views or to describe events in video data. Events can be expressed by interpreting collective behavior of physical objects over a certain period of time. In a simplistic manner, the behavior can be described by observing the total (or partial) duration during which an object appears in a given video clip. As mentioned earlier, the relative movement of an object with respect to other objects over the sequence of frames in which it appears is analyzed for event identification. For example, consider a user's query to search for the occurrence of a *slam-dunk* in a sports video clip. Modeling this particular event requires identification of at least two temporal *subevents* which include precise tracking of the motions of the player involved in the slam-dunk and the ball, especially when the ball approaches the hoop. The overall process of composing a slam-dunk event requires a priori specification of multiple temporal *subevents*. It may be noted that a simple temporal event can be expressed in form of a logical expression consisting of various spatial events that span a number of frames [13]. More complex temporal events can subsequently be defined recursively in terms of other temporal events.

Temporal intervals are extensively used for modeling temporal events [5]. Temporal intervals consist of time durations characterized by two endpoints, or *instants*. Intervals and instant-based representations are well studied topics [25]. A time instant is a zero-length moment in time, such as "*3:00 PM*." In contrast to time instants, time intervals are defined by two time instants representing their end points and their durations represent temporal intervals. The length of a temporal interval is identified by the difference of its endpoint values. The relative timing between two intervals can be determined from these endpoints. By specifying intervals with respect to each other rather than by using endpoints, we decouple the intervals from an absolute or instantaneous time reference, leading us to the binary temporal relations of Fig. 4. In the next section, we discuss the use of temporal intervals and their variations in modeling video data.

3.5.2 Temporal Interval-Based Video Modeling

In one of the approaches, both temporal and logical operators are used to develop spatio-temporal logic for specifying video semantics. In another approach, spatio-temporal operators with set-theoretic operators are used to specify video events in form of algebraic expressions. Such operations include merge, union, intersection, etc. The set-theoretic approach is also used for video production environments [15], [31], [43]. Several temporal interval based modeling approaches have been proposed in the literature. They include spatio-temporal logic, algebraic models, and hybrid temporal interval and trajectory based models. In the following paragraphs, we briefly discuss these approaches.

3.5.2.1 Spatio-Temporal Logic. In this approach, salient objects identified in a scene are represented by symbols, and scenes are represented by a sequence of state assertions capturing the geometric ordering relationships among the projections of the objects in that scene. The assertions specify the dynamic evolution of these projections in time as the objects move from frame to frame. The assertions are inductively combined through the Boolean connectives and temporal operators. Temporal and spatial operators, such as *temporal/spatial eventually* and *temporal/spatial always* are used for modeling video semantics in an efficient manner [6]. Fuzziness and incomplete specification of spatial relationships are handled by defining multilevel assertions that provide general to specific detail of event specifications [6].

Day et al. [13] use the notion of generalized temporal intervals initially proposed in [25] for temporal modeling of video data. The generalization is based on the binary temporal relations as shown in Fig. 4. A generalized relation, known as *n-ary* relation, is a permutation among *n* intervals, labeled 1 through *n*. The basis for the generalization is that two consecutive intervals satisfy the same temporal relation. The *n–ary* relations are used to build the video semantics in the form of a hierarchy. For this purpose, *simple temporal events* are first constructed from spatial events by using the *meets n-ary* temporal operator. The operands of this *meets* operator are the spatial event that is observed from frame to frame. In other words, the "persistence" of a specified spatial event over a sequence of frames is observed. This persistence eventually maps into a time interval corresponding to the duration of the persistence associated with the spatial event. The observed event is termed as a *simple temporal event*. Identification of *simple temporal events* requires evaluation of spatial and motion information of objects, captured in the VSDG model [13].

An example of a *simple temporal event* consisting of two spatial events is "*passing of a ball between two players*." This event can be characterized by relating two similar spatial events "*holding of the ball by player u*" and "*holding of the ball by player v*" with certain delay. Using the temporal operators of Fig. 4, simple temporal events can then be combined to build *composite temporal events*. "*Double pass*" is an example of a *composite temporal event* that can be captured recursively based on the simple temporal event. For specifying this event, we can use the *meet* operator joining two simple temporal events such as two single passes.

3.5.2.2 Algebraic Models. In this approach, temporal operators are used in conjunction with set operations to build formalisms that allow semantic modeling as well as editing capabilities for video data. For example, the framework discussed in [15] defines a set of algebraic operators to allow spatio-temporal modeling, as well as video editing capabilities. In this framework, temporal modeling is carried out by the spatio-temporal operators used in the spatio-temporal logic formalisms. These operators are implemented through functions that map objects and their trajectories into temporal events. Based on lisp-like operators for extracting items and lists, several functions are used to perform various video editing operations such as inserting video clips, and extracting video clips and images from other video clips.

Another algebraic video model based on hierarchical abstraction of *video expressions* representing scenes and events is presented in [43]. The model is used to provide indexing and content-based retrieval mechanisms. A video expression, in its simplest form, consists of a sequence of frames representing a meaningful scene. Compound video expressions are constructed from simpler ones through algebraic operations that include creation, composition, and description operators. Composition operators include several temporal and set operations. The set operators are used to generate complex video expressions, i.e., video segments, according to some desired semantics and description. Content-based retrieval of data is managed through annotating each video expression with field name and value pairs, defined by the user.

Algebraic modeling approach has been extended to develop oriented abstraction of video data [31]. A video object in this approach is identical to a video expression in [43] and corresponds to semantically meaningful scenes and events. An object hierarchy is built using IS–A generalizations and is defined on instances of objects rather than classes of objects. Such generalizations allow grouping of semantically identical video segments. Inheritance in such object hierarchy is based on interval inclusion, where some attribute/value pairs of a video object *A* are inherited by another video object *B* provided the raw video data of *B* is contained in that of *A*. Set operators supporting composition operations such as interval projection, merge, and overlap constructs, are used for editing video data and defining new instances of video objects.

3.5.2.3 Hybrid Temporal Interval and Trajectory-Based Models. Interval models described above have a major shortcoming: They do not support user's queries that include trajectory and motion sketches. In several applications, the complexity of semantics can force users to sketch the motion trajectories of objects to describe a scenario. Furthermore, motion of several objects over a period of time may constitute an event. A novel approach which combines trajectory information of multiple objects over time in terms of evolution of object motion is proposed in [12]. The approach uses a Petri-net based representation of binary temporal intervals of Fig. 4. A place in the Petri net represents the detailed 2D trajectory information of an individual object. The overall structure of the Petri net provides a higher level temporal semantics involving multiple object trajectories represented by its places. The hybrid approach provides a powerful formalism to describe events based on multiple objects and their motion. The recursive structure of Petri nets allow higher level semantic descriptions. For such a purpose, a place in the Petri net can represent a *subnet* from the lower level Petri net. A place can also describe events by textual annotation, rather than trajectory of motion. This formalism provides a succinct representation and storage of domain knowledge associated with video data. Fig. 6 shows an example scenario of a *touchdown* event from a football clip (http://www.nfl.com). In this clip, *Player 4* makes a touchdown pass while the defense player is blocked by *Player 63*. This involves *Player 4* and *Player 63* running to the right at the same time, followed by the ball being passed assuming a *parabolic trajectory*, and finally the

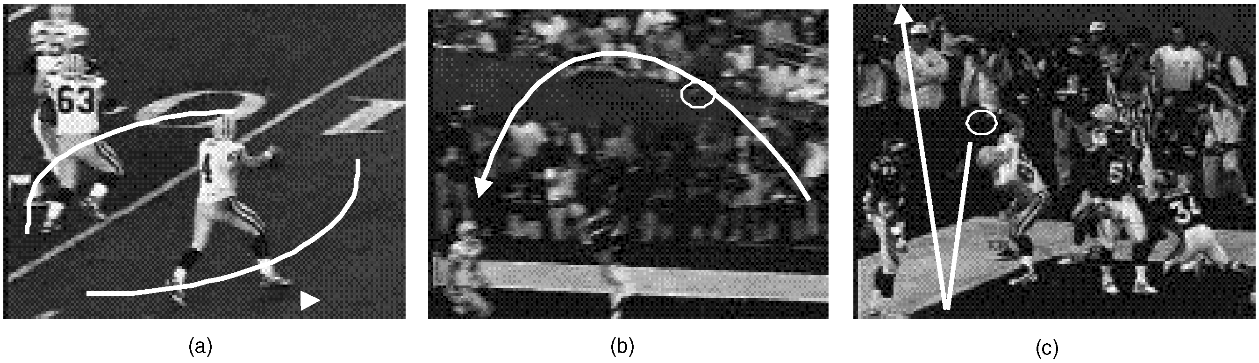


Fig. 6. An example video clip scenario for a *touchdown* play: (a) "Player 63 moving to right," represented by place P2, and "Player 4 moving to right," represented by place P1; (b) "ball passed," represented by place P3; (c) "ball slammed (score)," represented by place P4.

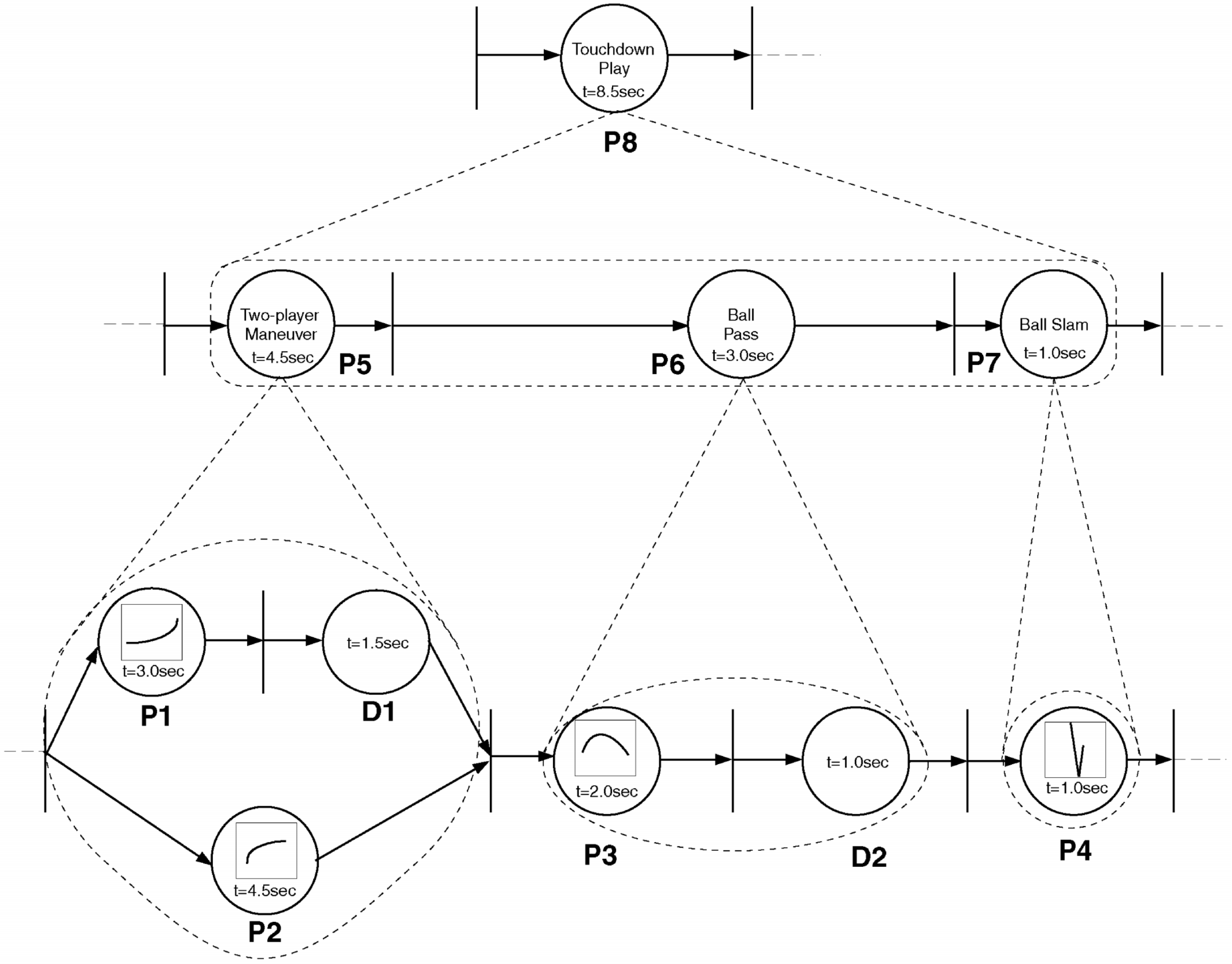


Fig. 7. Hybrid representation of the example *touchdown* play.

ball slammed on the ground. Fig. 7 shows the hierarchical representation using the hybrid model.

3.6 Summary and Challenges

Table 3 summarizes the important characteristics of the major video data modeling approaches surveyed in this paper. It can be noticed that most approaches that use an automatic mode of capturing video semantics cannot support high-level abstractions. This is due to the difficulty in capturing concepts that are difficult to map into a set of image and/or spatio-temporal features that can be automatically extracted from video data without human intervention. The use of the domain knowledge, such as the case in [39], probably is the only way by which higher level semantics can be incorporated into techniques that capture the semantics through automatic parsing. Also it can be observed from the table that the use of visual querying and browsing of video data is an important feature that must be provided as a part of the database. In particular, algebraic and logical expressions describing spatio-temporal semantics can pose difficulty in understanding and formulating queries. The approach in [6] uses a visual query facility, which provides an intuitive interface. In spatio-temporal modeling of video data, some degree of imprecision is intrinsic. To manage such imprecision, [6] employs different levels of precision in specifying spatial relationships among objects, unlike [13], which only supports the most precise and detailed representation.

The algebraic model presented in [15] has a limitation in the sense that it puts the burden on the user to define semantic functions related to video objects. Furthermore, these functions must be defined in terms of object trajectories. Two other algebraic approaches presented in [31], [43] require interactive formulation of video semantics by the user. These approaches provide flexibility in identifying the desired semantics. At the same time, they suffer from a shortcoming in terms of being unmanageable for naive users. Similarly, the high cost of human interaction can make these approaches impractical for large video databases.

In addition to the challenges we have listed for image databases, there are several data modeling issues specific to video data. In particular, semantic modeling of events,

TABLE 3
SURVEY OF DIFFERENT VIDEO DATABASE MODELS

Model	Spatial/Temporal Models (Event Representation)	Modeling Approach	Mode of Capturing Semantics	Query Specification
Smoliar et al. [39]	Predefined SCD-Based Model	Parsing/Segmentation	Automatic	Visual Browsing Tool
Yeung et al. [44]	Hierarchical Scene Transition Graph	Parsing/Segmentation	Automatic Semi Automatic	Visual Browsing Tool
Chen et al. [10]	–	Parsing/Segmentation	Automatic Semi Automatic	Visual Browsing Tool
Golshani et al. [15]	Algebraic	Object Identification & Motion Analysis	Automatic	Algebraic Expressions
Day et al. [13]	Spatio-Temporal Logic using objects & events	Object Identification & Motion Analysis	Manual	Logical Expressions
Bimbo et al. [6]	Spatio-Temporal Logic using objects & events	Object Identification & Motion Analysis	Semi Automatic	By Sketch
Oomoto et al. [31]	Algebraic using Video Objects	Semantic-Based Segmentation	Manual	Visual SQL-Based
Weiss et al. [43]	Algebraic using video expressions	Semantic-Based Segmentation	Manual	Algebraic Expressions

knowledge representation of spatio-temporal scenarios, and query formulation are the key issues needing extensive research. Temporal dimension of video data can introduce a high degree of imprecision and fuzziness. An interesting related problem is to classify domain-specific motion, and temporal events which can be used for evaluating the database performance [12].

4 MULTIMEDIA DOCUMENT MODELING

Another major issue that the database community needs to address is the management of multimedia documents. We believe, parallel to the explosive growth in computer and networking technologies, document repositories will soon become a reality and easy access to multimedia documents will make it essential to formally develop meta schema and indexing mechanisms for developing large scale multimedia document management systems.

An important issue for managing a large volume of multimedia documents is efficient indexing techniques to support querying of multimedia documents. Searching information about a document can be multidimensional. These include searching by spatio-temporal structures of documents, their logical organization, or by contents. For example, the query:

- *"Find documents that show a video clip of a basketball game, accompanied by textual information about other games' results,"*

requires searching documents by their spatio-temporal structures. Similarly, the query:

- *"Find documents that describe the assembly process of the transmission system of a car,"*

requires searching document databases by content. On the other hand, the query:

- *"Find all the other sections in this book that refers to the Image of the Himalayas of Chapter 7,"*

requires searching within a document based on its logical structure.

One of the major issues for multimedia document management systems is the integration of the data, that requires both temporal and spatial synchronizations of monomedia data to compose multimedia documents. In addition to this, logical organization of document components is desired to facilitate browsing and searching within and across documents. For managing documents, representation of composition and logical information in the form of a suitable meta schema is essential for designing efficient search strategies.

A generic architecture that highlights the overall process of document creation, management, and retrieval, is shown in Fig. 8. Our focus here is on the second layer of this architecture that deals with the composition and management aspects of multimedia documents.

Temporal synchronization is the process of coordinating the real-time presentation of multimedia information

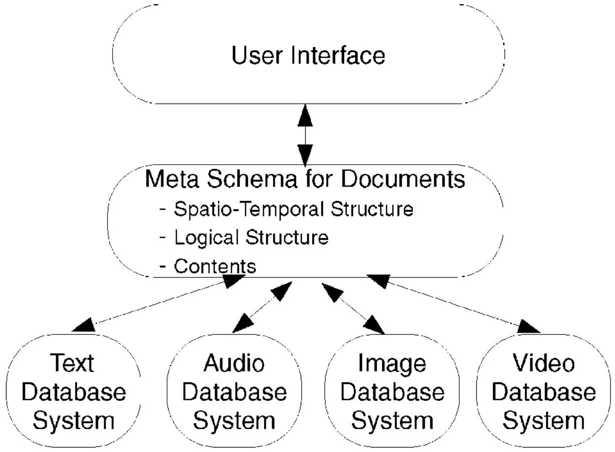


Fig. 8. A generic architecture for multimedia document management system.

and maintaining the time-ordered relations among component media. It is the process of ensuring each data element appears at the required time and is played out for a certain time period. A familiar example is the voice annotated slide show, where slides and voice data are played out concurrently.

Spatial composition describes the assembly process of multimedia objects on a display device at certain points in time. For text, graphics, image, and video, spatial composition includes overlay and mosaic, and requires processing such as scaling and cropping. For audio data, spatial operations include mixing of signals, gain, tone adjustment, and selectively playing out various audio signals on multichannel outputs (stereo quad, etc.).

In the following sections, we elaborate on two main aspects of document management; their spatio-temporal composition requirements and their organization models. Section 4.1 discusses the spatio-temporal modeling of documents. The organizational model is described in Section 4.2.

4.1 Composition Models for Multimedia Documents

In order to facilitate users to specify the spatio-temporal requirements, at the time of authoring a document, a composition model is needed. Recently, various such models have been proposed in the literature, which include language-based models, time-interval based models, and oriented models [1], [2], [18], [21], [23], [25], [32]. These models are described briefly in the following subsections.

4.1.1 Language Based Models

In this approach, a scripting language is used to describe the spatio-temporal structure of multimedia documents. The leading example is the HyTime model that uses SGML (Standard Generalized Markup Language). HyTime has been recognized as an ISO standard for multimedia document modeling in 1986 [1]. SGML has gained increasing popularity recently through the fame of its child, HTML, although it is a result of a decades-long effort. SGML basically defines a framework to describe the logical layout of the information in a structured format through a user-defined markup language. Defining metastructures involve location addressing of entities within data, querying of the structure and content of documents and most importantly, specification of measurement and scheduling of data contents along spatial and/or temporal axes. This last feature of the standard and the deserved popularity of markup schemes in data representation makes HyTime the ideal choice for multimedia document specification [1]. On the other hand, the multimedia technology still lacks “HyTime-aware” methodologies capable of creating and analyzing HyTime documents from the database management points of view.

A number of researchers have reported work involving SGML/HyTime structures [21], [32]. They mainly concentrate on document modeling and integrating HyTime-based information with databases. In their work, Özsu et al. describe a database application of SGML/HyTime documents for news-on-demand applications [32]. The documents follow a fixed logical structure and the document database is restricted to a certain schema. The document units are mapped into database objects in conformance with a predefined type hierarchy. Their work emphasizes the importance of spatial and temporal analysis and indexing of multimedia documents, but does not propose any approach to address this issue.

In [21], an alternate approach to the problem of storage and processing of structured documents within a DBMS framework, is presented. Realizing the advantages of a general purpose scheme, a document insertion mechanism using super *Document Type Descriptors* that allow handling of arbitrary documents in the database is presented [21]. Like the news-on-demand application in [32], the scheme uses an oriented DB manager called VODAK. Spatio-temporal indexing is explicitly referenced as an important research problem, although no specific results have been reported. However, content-based and general indexing is briefly mentioned.

In summary, the HyTime standard is expected to play a major role in leading the research activities in multimedia document modeling. However, the management aspects of HyTime-based documents in terms of searching and indexing are open research issues.

4.1.2 Interval-Based Models for Multimedia Documents

Recently, the use of Petri nets for developing conceptual models and browsing semantics of multimedia objects [18], [25] has been proposed. The basic idea in these models is to represent various components of multimedia objects as places and describe their inter-relations in the form of transitions. These models have been shown to be quite effective for specifying multimedia synchronization requirements and visualizing the composition structure of documents.

One such model is used to specify object level synchronization requirements which is both a graphical and mathematical modeling tool capable of representing temporal concurrency of media. In this approach Timed Petri Nets have been extended to develop a model that is known as Object Composition Petri Nets (OCPNs) [25]. The particularly interesting features of this model are the ability to explicitly capture all the necessary temporal relations. Each place in this Petri-net derivative represents the playout of a multimedia object while transitions represent synchronization points.

Several variations to the OCPN model have been proposed in the literature. One such variation deals with the spatial composition aspects of multimedia documents. For such composition, additional attributes are specified with each media place in the OCPN. These include: the size and location of the display area for different media within a document, a priority vector that describes the relative ordering among changing background/foreground locations of intersecting spaces for media display with time; an ordered list of unary operations, e.g., crop, scale, etc., applied to the data associated with the place, and a textual description about the contents of the media place.

As mentioned, the HyTime model suffers from a drawback and that is the extraction of various spatio-temporal and content semantics from this model can be quite cumbersome. On the other hand, the OCPN model not only allows extraction of the desired semantics and generation of a database schema, but also have the additional advantage of pictorially illustrating synchronization aspects of the information. In this regard this model is quite unique and, therefore, is also well suited for visual orchestration of multimedia document.

4.2 Organization Models for Multimedia Documents

From organizational structure point of view, a multimedia document can be viewed as a collection of related information objects, such as books, chapters, sections, etc. The logical structure of objects can be maintained in the form of a meta schema associated with each document. Meta information about such organization can be used for searching and accessing different parts of a document. Models for the logical structure of multimedia documents can be independent from the composition models. Such independence can support different presentation styles for a document that can be tailored to the target audience, as well as hardware display constraints.

The well-known organizational modeling paradigm of documents is based on hypermedia. There are basically three types of links used in a hypermedia environment. These include

- the base structure links for defining the organization of documents,
- the associative links for connecting concepts and accessing the same information from different contexts, and
- referential links that provide additional information on a concept within a document.

The HyTime model provides an elegant mechanism for the organizational structure of a document. Using SGML, a document's logical content is described by specifying the significant elements in that document along with the attributes associated with each such element, in a hierarchical manner. For example, an SGML specification of a textual report document may declare that it contains a title, an author and a body. Each of these elements would in turn have attributes specifying their structure.

The hypermedia-based multimedia document models have several attractive features. For example, they allow efficient path searching mechanisms for accessing information in various parts of the document [23]. Furthermore, they allow the development of oriented abstractions of documents. For this purpose, the document components are represented in form of a set of nodes which are related to each other through IS-A, IS-PART-OF, and AGGREGATE relationships. Associated with each node is a concept or a topic, and the semantic relationships among nodes are based on concepts. In other words, in this model, each node is an information unit, and oriented abstractions between two nodes can be represented using structural links.

Several hypermedia-based models of documents with object-oriented abstractions have been proposed in the literature [18], [21], [23], [32]. The model presented in [21], in essence, is HyTime model, as discussed earlier. Its hypermedia-based organization has been used to develop a multilayered architecture, known as VODAK. The layers consist of: a conceptual schemata level for accessing several multimedia databases, a second level that supports a document authoring environment by conceptualizing media objects, and a third level for the presentation of documents. The limitation in the design of VODAK system is that there is no explicit mechanism of querying based on contents associated with objects in a document.

Recently, the researchers in [23] proposed a hypermedia-based document model that uses the object-oriented paradigm. They describe a unique indexing scheme based on the underlying multistructure information of a document to optimize the indices and to provide efficient access to document elements. The document data model can be implemented using object-oriented technology. The model is augmented with an object-oriented query language syntax.

4.3 Summary

There is growing interest in multimedia document modeling and developing standards for document authoring. However, from database management point of view, very few results have been reported in the literature. We expect that the object-oriented technology can provide a powerful paradigm to meet the requirements of multimedia composition and organizational modeling.

5 Conclusion

In this paper, we have highlighted major technical issues and approaches in modeling and management of multimedia data, with emphasis on image, video and document data. We have tried to emphasize the vital role played by the image processing technology and knowledge representation in developing formalisms for multimedia data management. Overall, the approaches discussed in this paper provide a panoramic view spanning a variety of issues being researched by the multimedia database community, and give an idea of the scope and directions of future research in this important and promising field of endeavor. As our appreciation of underlying issues increases, and systems are refined, we may look forward to an exciting and productive age of multimedia databases and applications.

ACKNOWLEDGMENTS

This work has been partially funded by a National Science Foundation grant, under Grant No. IRI-9619812.

REFERENCES

[1] ISO/IEC 10744, *Information Technology—Hypermedia/Time-Based Structuring Language (HyTime)*, International Organization for Standardization, 1992.

[2] ISO 8613, *Information Processing—Text and Office Systems—Office Document Architecture (ODA) and Interchange Format*, International Organization for Standardization, 1993.

[3] A. Abella and J.R. Kender, "Qualitatively Describing Objects Using Spatial Prepositions," *Proc. 11th Nat'l Conf. Artificial Intelligence*, pp. 536–540, July 1993.

[4] A.A. Alatan, E. Tuncel, and L. Onural, "A Rule-Based Method for Object Segmentation in Video Sequences," *Proc. Int'l Conf. Image Processing*, vol. 2, pp. 522–525, Santa Barbara, Calif., Oct. 1997.

[5] J.F. Allen, "Maintaining Knowledge About Temporal Intervals," *Comm. ACM*, vol. 26, no. 11, pp. 832–843, Nov. 1983.

[6] A. Del Bimbo, E. Vicario, and D. Zingoni, "Symbolic Description and Visual Querying of Image Sequences Using Spatio-Temporal Logic," *IEEE Trans. Knowledge and Data Eng.*, vol. 7, no. 4, pp. 609–622, Aug. 1995.

[7] A. Brink, S. Marcus, and V. Subrahmanian, "Heterogeneous Multimedia Reasoning," *Computer*, vol. 28, no. 9, pp. 33–39, Sept. 1995.

[8] S.-K. Chang, Q.-Y. Shi, and C.-W. Yan, "Iconic Indexing By 2D Strings," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 9, no. 3, pp. 413–427, May 1987.

[9] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," *Proc. IEEE*, vol. 83, no. 5, pp. 705–741, May 1995.

[10] J.Y. Chen, C. Taskiran, E.J. Delp, and C.A. Bouman, "ViBE: A New Paradigm for Video Database Browsing and Search," *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries*, pp. 96–100, 1998.

[11] W.W. Chu, C.C. Hsu, A.F. Cardenas, and R.K. Taira, "Knowledge-Based Image Retrieval with Spatial and Temporal Constructs," *IEEE Trans. Knowledge and Data Eng.*, vol. 10, no. 6, pp. 872–888, 1998.

[12] S. Dagtas, W. Al-Khatib, A. Khokhar, and A. Ghafoor, "A Hybrid Content-Based Retrieval Approach for Video Data," Technical Report TR-ECE 98-13, Purdue Univ., Sept. 1998.

[13] Y.F. Day, S.D. Dagtas, M. Iino, A. Khokhar, and A. Ghafoor, "Spatio-Temporal Modeling of Video Data for On-Line Object-Oriented Query Processing," *Proc. IEEE Int'l Conf. Multimedia Computing and Systems*, pp. 98–105, May 1995.

[14] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The QBIC System," *Computer*, vol. 28, no. 9, pp. 23–32, Sept. 1995.

[15] F. Golshani and N. Dimitrova, "Retrieval and Delivery of Information in Multimedia Database Systems," *Information and Software Technology*, vol. 36, no. 4, pp. 235–242, May 1994.

[16] J. Hafner, H.S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient Color Histogram Indexing for Quadratic Form Distance Functions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 729–736, July 1995.

[17] C.C Hsu, W.W. Chu, and R.K. Taira, "A Knowledge-Based Approach for Retrieving Images by Content," *IEEE Trans. Knowledge and Data Eng.*, vol. 8, no. 4, pp. 522–532, Aug. 1996.

[18] M. Iino, Y.F. Day, and A. Ghafoor, "Spatio-Temporal Synchronization of Multimedia Information," *Proc. 1994 IEEE Int'l Conf. Multimedia Computing and Systems*, pp. 110–119, May 1994.

[19] A.K. Jain, Y. Zhong, and S. Lakshmanan, "Object Matching Using Deformable Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 3, pp. 267–278, Mar. 1996.

[20] R. Kasturi and R. Jain, "Dynamic Vision," R. Kasturi and R. Jain, eds., *Computer Vision*, pp. 469–480, IEEE CS Press, 1991.

[21] W. Klas, E.J. Neuhold, and M. Schrefl, "Using an Object-Oriented Approach to Model Multimedia Data," *Computer Comm.*, vol. 13, no. 4, pp. 204–216, May 1990.

[22] W. Krattenthaler, K.J. Mayer, and M. Zeiller, "Point Correlation: A Reduced-Cost Template Matching Technique," *Proc. ICIP*, pp. 208–212, 1994.

[23] K. Lee, Y.K. Lee, and P.B. Berra, "Management of Multi-Structured Hypermedia Documents: A Data Model, Query Language, and Indexing Scheme," *Multimedia Tools and Applications*, vol. 4, no. 2, pp. 199–223, Mar. 1997.

[24] J.H. Lim, H.H. Teh, H.C. Lui, and P.Z. Wang, "Stochastic Topology with Elastic Matching for Off-Line Handwritten Character Recognition," *Pattern Recognition Letters*, vol. 17, no. 2, pp. 149–154, Feb. 1996.

[25] T.D.C. Little and A. Ghafoor, "Interval-Based Conceptual Models for Time-Dependent Multimedia Data," *IEEE Trans. Knowledge and Data Eng.*, vol. 5, no. 4, pp. 551–563, Aug. 1993.

[26] B.M. Mehtre, M.S. Kankanhalli, A.D. Narasimhalu, and G.C. Man, "Color Matching for Image Retrieval," *Pattern Recognition Letters*, vol. 16, no. 3, pp. 325–331, Mar. 1995.

[27] J. Meng, Y. Juan, and S.-F. Chang, "Scene Change Detection in a MPEG Compressed Video Sequence," *Proc. SPIE*, vol. 2,419, pp. 14–25, 1995.

[28] M. Misra and V.K. Prasanna, "Parallel Computations of Wavelet Transforms," *Proc. Int'l Conf. Pattern Recognition*, Sept. 1992.

[29] A. Nagasaka and Y. Tanaka, "Automatic Video Indexing and Full Video Search for Object Appearances," *Proc. Second Working Conf. Visual Database Systems*, pp. 119–133, IFIP WG 2.6, Oct. 1991.

[30] V.E. Ogle and M. Stonebraker, "Chabot: Retrieval from a Relational Database of Images," *Computer*, vol. 28, no. 9, pp. 40–48, Sept. 1995.

[31] E. Oomoto and K. Tanaka, "Ovid: Design and Implementation of a Video-Object Database System," *IEEE Trans. Knowledge and Data Eng.*, vol. 5, no. 4, pp. 629–643, Aug. 1993.

[32] M.T. Ozsu, D. Szafron, G. El-Medani, and C. Vittal, "An Object-Oriented Multimedia Database System for a News-On-Demand Application," *ACM Multimedia Systems J.*, vol. 3, nos. 5-6, pp. 182–203, Nov. 1995.

[33] S. Ravela, R. Manmatha, and E.M. Riseman, "Image Retrieval Using Scale-Space Matching," *Proc. Fourth European Conf. Computer Vision*, pp. 273–282, 1996.

[34] E. Remias, G. Sheikholeslami, A. Zhang, and F. Syeda-Mahmood, "Supporting Content-Based Retrieval in Large Image Database Systems," *Multimedia Tools and Applications*, vol. 4, no. 2, pp. 153–170, Mar. 1997.

[35] L.A. Rowe, J. Boreczky, and C. Eads, "Indexes for User Access to Large Video Databases," *Proc. IS and T/SPIE Int'l Symp. Electronic Imaging: Science and Technology*, San Jose, Calif., Feb. 1994.

[36] T. Sakai, M. Nagao, and S. Fujibayashi, "Line Extraction and Pattern Detection in a Photograph," *Pattern Recognition*, vol. 1, no. 3, pp. 233–248, Mar. 1969.

[37] H. Samet and A. Soffer, "MARCO: Map Retrieval by Content," *IEEE Trans. Pattern Analysis and Machine*, vol. 18, no. 8, pp. 783–798, Aug. 1886.

[38] S. Santini and R. Jain, "Similarity Queries in Image Databases," *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition*, pp. 646–651, 1996.

[39] S.W. Smoliar and H. Zhang, "Content-Based Video Indexing and Retrieval," *IEEE Multimedia*, vol. 1, no. 2, pp. 62–74, Summer 1994.

[40] R.K. Srihari, "Automatic Indexing and Content-Based Retrieval of Captioned Images," *Computer*, vol. 28, no. 9, pp. 49–56, Sept. 1995.

[41] M.J. Swain and D.H. Ballard, "Indexing Via Color Histograms," *Proc. Third Int'l Conf. Computer Vision*, pp. 390–393, 1990.

[42] S.L. Tanimoto, *Elements of Artificial Intelligence Using Common LISP*, Computer Science Press, 1990.

[43] R. Weiss, A. Duda, and D.K. Gifford, "Composition and Search with a Video Algebra," *IEEE Multimedia*, vol. 2, no. 1, pp. 12–25, Spring 1995.

[44] M.M. Yeung, B.-L. Yeo, W. Wolf, and B. Liu, "Video Browsing Using Clustering and Scene Transitions on Compressed Sequences," *Proc. IS and T/SPIE Multimedia Computing and Networking*, 1995.

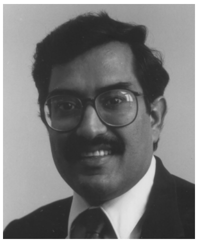
[45] A. Yoshitaka, S. Kishida, M. Hirakawa, and T. Ichikawa, "Knowledge-Assisted Content-Based Retrieval for Multimedia Database," *IEEE Multimedia*, vol. 1, no. 4, pp. 12–21, Winter 1994.

[46] H.J. Zhang, A. Kankanhalli, and S.W. Smoliar, "Automatic Partitioning of Full-Motion Video," *Multimedia Systems*, vol. 1, no. 1, pp. 10–28, 1993.

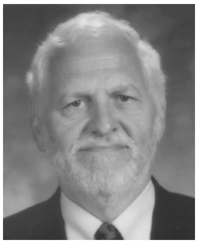
[47] H.J. Zhang, C.Y. Low, Y. Gong, and S.W. Smoliar, "Video Parsing Using Compressed Data," *Proc. SPIE*, vol. 2,182, pp. 142–149, 1994.

Wasfi Al-Khatib received his BS degree in computer science from Kuwait University in 1990, and his MS degree in computer science from Purdue University in 1995. He is currently a PhD candidate in the School of Electrical and Computer Engineering at Purdue University. His research interests include multimedia information systems, artificial intelligence, and software engineering. He is a member of the ACM, the UPE, and the IEEE Computer Society.

Y. Francis Day received his BS degree in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1989; his MS degree in computer engineering from Pennsylvania State University, University Park, in 1991; and his PhD in electrical and computer engineering from Purdue University in 1996. He is currently with the Multimedia Documentation Program at Siemens Corporate Research (SCR) in Princeton, New Jersey, where he has been a member of the technical staff since May 1997. Before joining SCR, he worked from 1996 to 1997 for the Network Management Group of Alcatel Data Networks in Ashburn, Virginia. His current research interests include interactive multimedia information systems, video data modeling, distributed/networked multimedia documentation systems based on SGML/XML, and multimedia document management systems. He is a member of the IEEE, the IEEE Computer Society, and the ACM.



Arif Ghafoor received his BS degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 1976; and the MS, MPhil, and PhD degrees from Columbia University in 1977, 1980, and 1984, respectively. In the Spring of 1991, he joined the faculty of the School of Electrical Engineering at Purdue University, where he is now an associate professor. Prior to joining Purdue University, he was on the faculty of Syracuse University from 1984 to 1991. His research interests include design and analysis of parallel and distributed systems, and multimedia information systems. He has published in excess of 100 technical papers in these areas. Currently, he is directing a research laboratory in distributed multimedia systems at Purdue University. His research in these areas has been funded by DARPA, NSF, NYNEX, AT&T, Intel, IBM, Fuji Electric Corporation, and GE. He has served on the program committees of various IEEE and ACM conferences. Currently, he is on the Editorial Board of *ACM Multimedia Systems*, the *Journal of Multimedia Tools and Applications*, and the *Journal of Parallel and Distributed Databases*. He is, or was, a guest/co-guest editor of special issues of several journals, including *IEEE JSAC*, *ACM Multimedia Systems*, the *Journal of Multimedia Tools and Applications*, and the November/December 1998 and May/June 1999 special sections of *IEEE Transactions on Knowledge and Data Engineering*. He is a fellow of the IEEE and a member of the IEEE Computer Society.



P. Bruce Berra received the BS and MS degrees from the University of Michigan, and the PhD degree from Purdue University. He is now director of the Information Technology Research Institute in the College of Engineering and Computer Science and Regents Professor of Information Technology in the Department of Computer Science and Engineering at Wright State University in Dayton, Ohio. He is also an emeritus and research professor of electrical engineering and computer science at Syracuse University. Prior to July 1996, he was director of the New York State Center for Advanced Technology in Computer Applications and Software Engineering (CASE), and a professor of electrical and computer engineering and a faculty member in computer and information science at Syracuse University. He previously served as chair of industrial engineering and operations research at Syracuse University, and has taught at the University of Michigan's Dearborn campus, Boston University, and Purdue University. His industrial experience includes periods of service with Hughes, Bendix, and IBM. He is currently president of PBB Systems, a knowledge and database consulting firm. He has served on the IEEE Computer Society Board of Governors; was editor-in-chief of the IEEE Computer Society Press; was on the IEEE Computer Society Press Editorial Board and a member of the Computer Society Distinguished Visitors Program; and was general chair and program chair of the IEEE International Conference on Data Engineering. He has served as an editor of *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transaction on Software Engineering*, *IEEE Transactions on Computers*; and has chaired the IEEE Computer Society Transactions Advisory Committee. He is one of the three co-founders of *IEEE Transactions on Knowledge and Data Engineering.* He pursues research interests in multimedia information systems, parallel processing for very large data and knowledge bases, and optical database machines. He was recently presented the L.C. Smith College of Engineering and Computer Science Award for Excellence in Scholarship for his contributions to knowledge in his field. In 1994 and 1995, he was a finalist in the Supporter of Entrepreneurship category of the Entrepreneur of the Year Institute Awards. He is a fellow of the IEEE, a member of the IEEE Computer Society, and co-guest editor for the November/December 1998 and May/June 1999 special sections of *IEEE Transactions on Knowledge and Data Engineering*.