Metadata in Video Databases

Ramesh Jain
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
email: jain@ece.ucsd.edu
and
Arun Hampapur
1101 Beal Ave
Department of Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI 48109
email: arun@eecs.umich.edu

Abstract

Video is composed of audio-visual information. Providing content based access to video data is essential for the sucessful integration of video into computers. Organizing video for content based access requires the use of video metadata. This paper explores the nature video metadata. A data model for video databases is presented based on a study of the applications of video, the nature of video retrieval requests, and the features of video. The data model is used in the architectural framework of a video database. The current state of technology in video databases is summarized and research issues are highlighted.

1 Introduction

The invention of printing caused the first revolution in human communication [18]. The second revolution in communication is underway with digital video [8] The progress of this revolution is critically dependent on the ability of computer systems to provide content based access to digital video. Video databases fill in this critical role of storage and retrieval mechanisms. Video data unlike text is opaque to computers. This demands that the video database posses some representation in addition to the video which allows the content based access of video. This paper presents the essential characteristics of the metadata necessary to provide content based video access.

Research in the area of digital video analysis and video databases has been focussed on the fine grain manipulation of video [2, 5, 15, 19, 21]. Some current research is addressing the problem of video retrieval at a slightly coarser grain. Davenport et al [7] and Davis [9] address the problem of manual video annotation and the types of annotation symbols needed. Hampapur [14] address the problem of video indexing based on different types of video search keys.

Current research efforts have been motivated from the point of view of video composition and multimedia authoring, resulting in the problem of video access being posed at a level of shots and scenes. From the consumer perspective video access will be at a much coarser level. For example, the home video user is more likely to request a particular type of program as opposed to requesting a particular kind of shot. The discussion in this paper considers both the consumer and producer perspectives providing a balanced view of the problem. The problem of video storage and retrieval is addressed over a wide range of granularities.

A set of terms used in the paper are defined in Section 2. Section 3 presents a classification of video based on usage. Section 4 presents a classification of typical queries in a video database. The discussion on applications and queries is used to motivate the design of a data model for video and its use in a typical video database architecture in Section 5. Research issues and directions are discussed in Section 6. The paper concludes with a summary in Section 7.

2 Terminology

The following are terms which are used in the paper:

Digital Video: Any audio visual temporally presentable information stored in a digital format.

Meta Data: Data about data [12]. In the context of video meta data refers to data which is used in organising video to facilitate content based retrieval.

Features: A set of numbers and symbols resulting from applying transformations to the audio visual data of a video.

3 Applications of Video

Video is used in many different types of applications to capture information. The nature of the information captured and the purpose for which the video is used greatly affect the structure of the video. This section groups different applications into categories based on the *purpose* for which video is used. The nature of video in each of these groups is enunciated and used as a guide to the design of the video database system. A detailed study of classification of video programs can be found in [11]. Techniques for analyzing and representing the content of communications have been studied in [17]. The following is a list of purpose based classes. Examples of applications in these classes are listed in table 1. A more detailed study of applications can be found in [14].

Entertainment: The information presented in this class of videos is highly stylized depending on the particular sub-category. Detailed classification and content description of current cable programming can be found in [1].

Information: The purpose of videos here is to convey information to the viewer.

Communications: The purpose of videos in this category are communication of information.This is not yet wide spread but is expected to grow significantly in the future [10].

Data Analysis: A number of different scientific disciplines use video for recording data during experiments and using the video as a source for further data analysis. In these applications the amount of time required to analyze the data tends to be very large compared to the data collection time. This is due to the lack of proper access and analysis tools for dealing with video. Digital video and video database technology are expected to exponentially increase the use of video as a medium for data capture and analysis. Some examples of video as an analysis tool can be found in [20].

Category	Sub-Category	Examples
Entertainment	Fiction	Films, TV Shows Music videos
	Non-Fiction	Sports, Cop shows Nature, Adventure Talk shows
	Interactive	Game shows Shopping shows
Information		News, Education Training Shows
Communication		Video Conferences Presentations
Data Analysis		Sports, Medicine Surveillance

Table 1: Applications of Video

4 Nature of Queries in Video Databases

The design of any database system is driven by the nature of the queries that will be made about the data. This section studies queries expected in video databases based on their common characteristics. The criteria used for grouping are the *certainty of a query*, *the dimensionality of the query* and *data dependence*.

4.1 Query Certainty

The certainty of a query can be specified in terms of the type of matching operator used to satisfy the query.

Exact Matches: A query is certain if it is of the form *Find me all records with* **field-#1 = X** where **X** belongs to some well defined set of labels and the operation of equality is well defined. For example, *Retrieve all CNN reports with < subject = Simpson Trial>*

Inexact Matches: In the case of data derived features the feature sets tend to be large, and the concept of equality is not well defined. Here equality translates to similarity, where similarity is defined in terms of some distance measure and a set of bounds on the distance measure. For example, **Retrieve all CNN reports with forest *like* scenes.** Here *like* denotes the nature of the matching operation. Inexact matches are typically used in systems with query by example capability.

4.2 Query Dimensionality

The information required to satisfy a query can be derived from different aspects of video. The minimum number of aspects of video that have to be considered in order to satisfy a query can be called its dimension. The following is a brief description of the different dimensions of a query.

Audio: The query depends only on the audio track. Such queries will be predominant in applications like retrieval of lecture, or video conference videos, where most of the information is in the audio component. For example, *Retrieve video with < discussion = budget>*

Visual: The response to these queries requires the use of visual information from the video. Visual queries can further be classified into the following:

- **Spatial:** These queries rely solely on the spatial information in the video. Example *Retrieve clips with < background = capitol>*
- **Temporal:** These are queries which rely only on the temporal information in video. For example, *Retrieve clips with < camera pan from left to right>*
- **Spatio-Temporal:** These are queries which rely on the spatio temporal aspects of video. *Retrieve clips with < a car chase >*

Audio-Visual: These queries require the use of information from all the aspects of video. Such queries tend to be more abstract in nature. For example, *Find programs with < comical treatment of the government>*

In many cases the queries will tend to be some weighted combination of the above dimensions of the video. This categorization of queries is used in the design of the data model for video.

4.3 Data Dependence

The queries discussed in the previous sections dealt with the content of video. There are many queries which can be satisfied without a content analysis of the video. Statistical querries are a typical example, *How many films in database with < director = spielburg>*

5 The Video Database System

This section presents the system architecture for a video database. This architecture (figure 2) is an adaptation of the Xenomania System Architecture [6]. Xenomania was used for the face image data management. Figure 3 shows the structure of the *interactive video processor*. The video database uses a data model called **ViMod**.

5.1 Video Data model: ViMod

This section proposes a data model for video databases. This data model is called **ViMod**. ViMod is designed to represent spatio-temporal information. The spatial representation uses a modified version of the **VIMSYS** model [13]. The ViMod data model takes into account the various aspects of video and various usage scenarios. The design goals for ViMod data model are summarized below:

Temporal Media: Video is a temporal medium. It captures and presents information over time. This data model must account for the *temporal nature of video.*

Application Purpose: The different kinds of applications of video were presented in section 1. The data model should accommodate data from all *application types.*

Reuse: Video data stored in a database may be generated for a certain purpose, but retrieved and used for an unrelated context. The data model should allow for such *reuse of video.*

Query Types: The data model should be able to support both exact match queries and similarity based queries.

The basic representational unit for video in ViMod is a *temporal interval*: a finite closed interval in the physical time required to present a video (viewing time)[16]. Thus any data request query to a video database results in set of temporal intervals.

5.1.1 Characteristics of Video Features

Given any *temporal interval* of video several different types of features of the video can be represented or highlighted during that interval. The following is a list of characteristics of these features.

Content Dependence: A feature is said to be content independent if the feature is not directly available from the video data. For example *the budget of a video* is not normally available in the contents of a video, where as *the story* can always be understood by viewing the video. Content dependent features are called *data features* and content independent features are called *meta features.*

Temporal Extent: Certain aspects of video can be specified based on viewing a single frame in a temporal interval (for example, the dominant color in the video, spatial locations of objects, object labels etc), whereas other features like motion can be specified only based on a time interval (like feature track, type of action, etc). Temporally extended features are called *video features*, others are referred to as *image features*.

Labeling: The changes that occur in video can be tracked (for example the track of a basketball in game footage) over the extent of a time interval. Domain models are used to assign a *qualitative label* to the feature (in the case of a basketball, labels like, *pass, dribble, dunk* etc). Domain model based labels of video are refered to as *qualitative feature (Q-Feature)*. Features which rely on low level domain independent models like object trajectories are called *raw feature (R-Feature)*.

5.1.2 Feature Classes in Video

Figure 1 shows the structure of the ViMod data model for representing video. The model has five classes of features. Each of the features applies to a *temporal interval*. The boxes in the figure partition the feature space based on *characteristics of the features*. The following is a brief description of the feature classes with an associated table which gives the typical features in each feature class and their typical values.

Video Q-Features: These are content dependent, temporally extended, labeled features of video (table 2).

Feature	Typical Value
Shot Distance	Long, Medium, Close up
Shot Angle	Low, Eye Level, High
Shot Motion	Tracking, Dolly, Pan
Shot Objects	One Shot, Two Shot
Audio Labels	Dialogue, Music, SFX
Shot Props	Color, Texture, lighting
Time Frame	What point in history
Video Class	News, Sports
Video Type	has objects, no objects
Point of View	Whose point of view
Subjective Properties	Intentions, Emotions
Object Properties	People, Trees

Table 2: Video Q-Features of a Time Interval

Video R-Features: These are content dependent, temporally extended, raw data values (table 3). It is important to note that the values that the feature takes is a set which is indexed by time.

Feature	Typical Value
Object Track	Set of image positions
Camera Pan	Camera Pan in degrees
Camera Tilt	Camera Tilt in degrees
Camera Height	Camera Height in meters
Audio Levels	Dialogue dB levels
Lighting levels	Average Lux

Table 3: Video R-Features of a Time Interval

Image Q-Features: These are content dependent, single frame, labeled features of video (table 4).

Feature	Typical Value
Image Brightness	Indoor, Outdoor, Cloudy
Texture Type	Random, Regular, Oriented
Ambient Lighting	Blue, Red
Spatial Activity	Clustered,Uniform
Audio Properties	Pitch, Loudness, Timbre
Object Name	Tree, Car
Object Color	Green, Cyan
Object Location	Center, Left, Right
Object Structure	Shape and Size
Audio Keywords	Begin, End

Table 4: Image Q-Features of a Time Interval

Image R-Features: These are content dependent, single frame, raw feature values of images (table 5).

Feature	Typical Value
Histograms	Float Arrays
Gradient Maps	Image
Edge Maps	Image
Feature Maps	Image
Audio FFT Maps	Float Arrays

Table 5: Image Features of a Time Interval

Meta Features: These are content independent features of video. They in general apply to a complete video and rarely to smaller time intervals. Such features are referred to as *meta features* (table 6).

Feature	Typical Value
Producer Info	Name
Date of Production	Date
Length	Number of Frames
Original Medium	Film, video

Table 6: Meta Features of Video

5.2 System Architecture

The architecture of the complete video database system is shown in figure 2. The user interacts with the system through a *graphical user interface*. The user interface has direct access to the video data. The insertion and retrieval of video data is handled by the *interactive video processor*, which uses a number of video processing routines to generate instances of the data model during the insertion process. The interactive video processor also handles the reformulation of the user queries. The *database* maintains the ordering in the video data and the links to the physical storage.

5.3 Interactive Video Processor

The interactive video processor has two parts, the *insertion module* which allows the user to interactively specify various aspects of the data model during the insertion process and the *query formulation unit* which allows the user to specify queries during the retrieval process. The insertion module has three interactive units

Video Feature Extractor: This unit interactively extracts raw feature vectors from the video.

Video Classifier: This unit uses the raw feature vectors along with a set of domain models to generate qualitatively labeled features.

Image Feature Extractor: Interactive extraction of image based features like regions, lines, etc are handled by this unit.

Image Classifier: The image features are labeled qualitatively based on some set of domain models.

Annotator: This allows the user to annotate the video a set of labels.

Object Linker: In the ViMod data model each class of features applies to a *time interval*. Given a video, a number of different sets of time intervals will be generated. The object linker performs the function of maintaining the temporal relationship between the different time intervals. A detailed study of time intervals and their relationships can be found in [4].

6 Research Directions

This paper has outlined the various aspects of video metadata necessary for providing content based video access. There are several very interesting and important problems to be addressed in designing video database systems. Several of the issues in video databases are similar to issues in spatio-temporal databases [3]. The following is a list of the major research problems to be addressed specifically in the context of video databases.

Program Categorization: What is good way of classifying videos to achieve the best retrieval results.

Subject Indexing Terms: What is a good set of subject index terms to use for describe the content of video.

Low Level Index Terms: What is a good set of terms for describing the cinematographic properties of video.

Query Specification: The issues of providing the correct query language which allows easy formulation of queries.

User Interface: Since a video database deals with spatio-temporal information, the issue appropriately presenting the information to the user is very important.

Interactive Data Analysis: The design and effective integration of interactive data analysis techniques with user input to improve the overall system performance.

Performance Measures: These are essential to evaluate the efficacy of a set of features in a video database system.

Implementation Issues: Representations for video data and efficient means of implementing the different functionalities required in video databases.

6.1 Future Automation Trends

Current techniques in video processing can automatically partition video into segments based on camera motion breaks. A few techniques can also deal with special transition effects [15]. Video labeling based on low level video properties is currently a research problem. The next five years will see the maturing of syntactic and low level video processing techniques mature and become commercially applicable.

Semantic or high level video annotation and processing is entirely manual at this time. This technology is likely to be partially automated over the long run, but will remain interactive till significant break through is made in the fields of computer vision and artificial intelligence. Video annotation is at present a post production process, where the annotation is done after the video is produced. The annotation point will gradually advance towards the production point in the future.

7 Summary

This paper has presented the metadata issues for content based video access. The paper has adopted the video database system design perspective. An architecture for a video database has been presented which uses the new data model for video, **ViMod**. The design of the model was based on the study of the meta characteristics of queries and video features. Current research topics, future research directions and the expected progress of video database technology and its direction were discussed.

References

[1] *Broadcasting and Cable Yearbook 1993.* Broadcasting and R R Bowker, 1993.

[2] Akihito Akutsu, Yoshinobu Tonomura, Hideo Hashimoto, and Yuji Ohba. Video indexing using motion vectors. In *Proceedings of SPIE: Visual Communications and Image Processing 92*, November 1992.

[3] Khaled K Al-Taha, Richard T Snodgrass, and Michael D Soo. Bibliography on spatiotemporal databases. *SIGMOD Record*, 22(1):59–67, March 1993.

[4] James F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, November 1983.

[5] Farshid Arman, Arding Hsu, and Ming-Yee Chiu. Image processing on compressed data for large video databases. In *Proceedings of the ACM MultiMedia*, pages 267–272, California, USA, June 1993. Association of Computing Machinery.

[6] Jeffrey Bach, Santanu Paul, and Ramesh Jain. An interactive image management system for face information retrieval. *IEEE Transaction on Knowledge and Data Engineering, Special Section on Multimedia Information Systems*, June 1992.

[7] Gloriana Davenport, Thomas Aguirre Smith, and Natalio Pincever. Cinematic primitives for multimedia. *IEEE Computer Graphics & Applications*, pages 67–74, July 1991.

[8] Duncan Davies, Diana Bathurst, and Robin Bathurst. *The Telling Image: The Changing Balance between Pictures and Words in a Technological Age.* Clarendon Press, 1990.

[9] Marc Davis. Media streams: An iconic visual language for video annotation. In *IEEE Symposium on Visual Languages*, pages 196–202. IEEE Computer Society, 1993.

[10] Edward M Dickson and Raymond Bowers. *The Video Telephone: Impact of a New Era in Telecommunications.* Praeger Publishers, 1973.

[11] P Gould, J Johnson, and G Chapman. *The Strucutre of Television.* Pion Ltd, 1984.

[12] DAFTG: Database System Study Group. Reference model for DBMS standardization. *SIGMOD Record*, 15(1):19–58, March 1986.

[13] Amarnath Gupta, Terry Weymouth, and Ramesh Jain. Semantic queries with pictures: the VIMSYS model. In *Proceedings of the 17th International Conference on Very Large Data Bases*, September 1991.

[14] Arun Hampapur. *Digital Video Indexing in Video Databases (*in preparation*).* PhD thesis, The University of Michigan, 1994.

[15] Arun Hampapur, Ramesh Jain, and Terry Weymouth. Digital video segmentation. In *Prooceedings Second Annual ACM MultiMedia Conference and Exposition.* Association of Computing Machinery, October 1994.

[16] Ira Konigsberg. *The Complete Film Dictionary.* Penguin Books, 1989.

[17] Klaus Krippendorf. *Content Analysis An Introduction to Its Methodology*. Sage Publications, 1980.

[18] Marshall McLuhan. *The Guttenberg Galaxy: the making of typographic man*. University of Toronto Press, 1962.

[19] Akio Nagasaka and Yuzuru Tanaka. Automatic video indexing and full-video search for object appearances. In *2nd Working Conference on Visual Database Systems*, pages 119–133, Budapest, Hungary, October 1991. IFIP WG 2.6.

[20] John W Northrip, Gene A Logan, and Wayne C McKinney. *Introduction to Biomechanic Analysis of Sport*. WM.C. Brown Company, 1974.

[21] H J Zhang, A Kankanhalli, and S W Smoliar. Automatic partitioning of video. *Multimedia Systems*, 1(1):10–28, 1993.

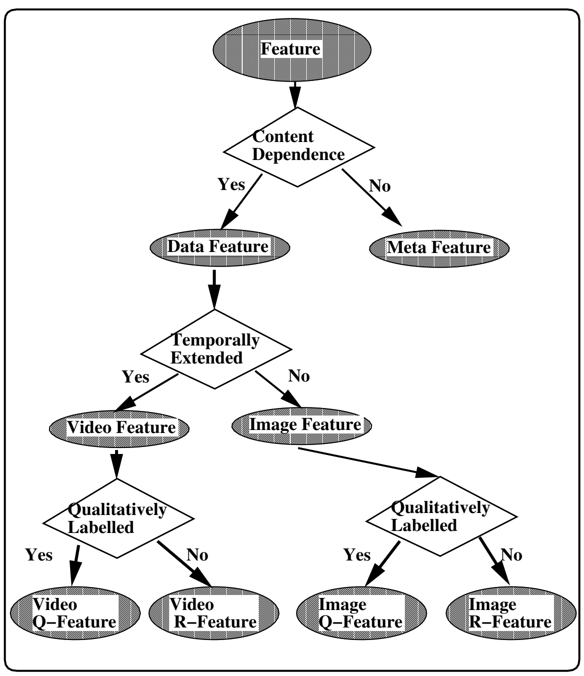

Figure 1: The **ViMod**: Video Data Model

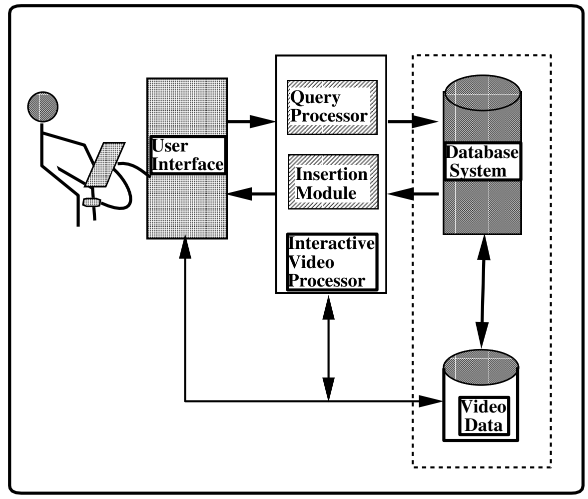

Figure 2: The Video Database System Architecture

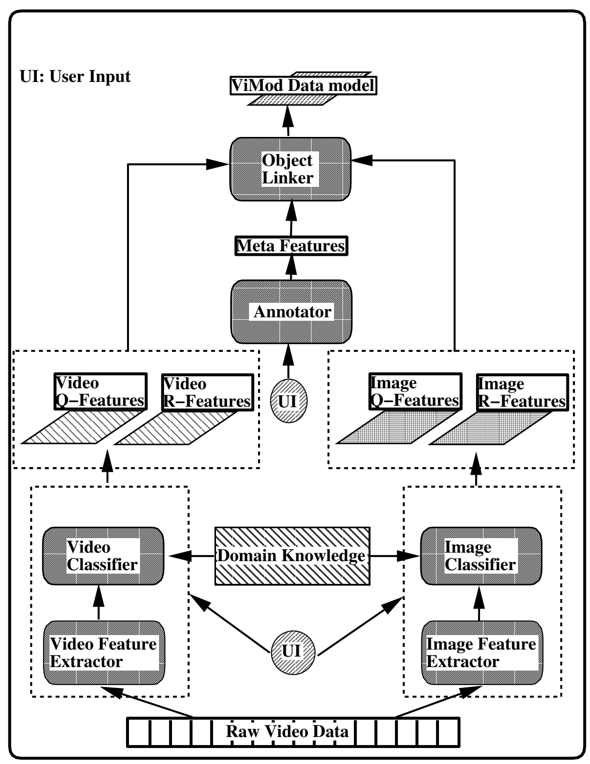

Figure 3: The Interactive Video Processor