ATTACHMENT C

TO REQUEST FOR *EX PARTE* REEXAMINATION OF U.S. PATENT NO. 7,932,923

Spatio-Temporal Modeling of Video Data for On-Line Object-Oriented Query Processing

Young Francis Day, Serhan Dağtaş, Mitsutoshi Iino, Ashfaq Khokhar, and Arif Ghafoor
Distributed Multimedia Systems Laboratory
School of Electrical Engineering
Purdue University, West Lafayette, IN 47907

Abstract

This paper presents a framework for data modeling and semantic abstraction of image/video data. The framework is based on spatio-temporal information associated with salient objects in an image or in a sequence of video frames and on a set of generalized n-ary operators defined to specify spatial and temporal relationships of objects present in the data. The methodology presented in this paper can manifest itself effectively in conceptualizing events and heterogeneous views in multimedia data as perceived by individual users. The proposed paradigm induces a multilevel indexing and searching mechanism that models information at various levels of granularity and hence allows processing of content-based queries in real time. We also devise a unified object-oriented interface for users with heterogeneous views to specify queries on the unbiased encoded data. Currently this framework is being developed to realize a highly integrated multimedia database architecture. [1]

1 Introduction

Recent advances in broadband networking, high performance computing, and storage systems have resulted in a tremendous interest in digitizing large archives of multimedia data and providing interactive access to users. Many future multimedia applications will require retrieval of video data including searching, browsing, selective replays, editing, etc. Due to the shear volume of the data, all these capabilities require efficient computer vision/image processing algorithms for automatic indexing and abstraction of video data. Subsequently, powerful indexing and data retrieval techniques need to be employed to support content-based query processing.

The key characteristic of video data is the spatial/temporal semantics associated with it, making video data quite different from other types of data such as text, voice and images. A user of video database can generate queries containing both temporal and spatial concepts. However, considerable semantic heterogeneity may exist among users of such data due to difference in their pre-conceived interpretation or intended use of the information given in a video clip. Semantic heterogeneity has been a difficult problem for conventional database [6], and even today this problem is not clearly understood. Consequently, providing a comprehensive interpretation of video data is a much more complex problem.

Most of the existing video database systems either employ image processing techniques for indexing of video data [5, 3, 10, 2] or use traditional database approaches based on keywords or annotated textual descriptors [11, 15]. However, most of these systems lack the ability to provide a general-purpose, automatic indexing mechanism which renders an unbiased description of video data. Also they do not handle the semantic heterogeneity efficiently. In order to address the issues related to user-independent view and semantic heterogeneity, we propose a framework for semantically unbiased abstraction of video data. The framework is based on spatio-temporal information associated with salient objects in an image or in a sequence of video frames and on a set of generalized n-ary operators defined to specify spatial and temporal relationships of objects present in the data. The methodology presented in this paper can manifest itself effectively in conceptualizing events and heterogeneous views in multimedia data as perceived by individual users. The proposed paradigm induces a multilevel indexing and searching mechanism that models information at various levels of granularity and hence

[1] This research was supported in part by the National Science Foundation under grant number 9418 767-EEC and in part by ARPA under contract DABT63-92-C-0022ONR.

0-8186-7105-X/95 $4.00 © 1995 IEEE

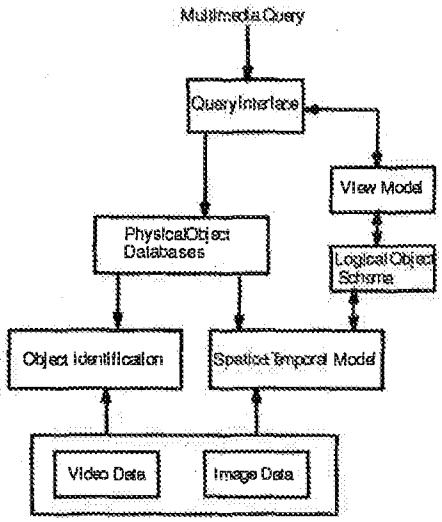

Figure 1: System abstraction

allows processing of content-based queries in real time. However, a unified framework is needed for the users to express and for the system to process semantically heterogeneous queries on the encoded data. For this purpose, we propose an object-oriented interface that provides an elegant paradigm for representing heterogeneous views of the users. The architecture of the proposed system is shown in Figure 1.

The organization of this paper is as follows. Section 2 presents the framework for characterizing various events in video data. A video database architecture based on that framework is proposed in Section 3. In Section 4, an object-oriented approach is presented for users to specify the perceived view of video data. The paper is concluded in Section 5.

2 Framework for Characterizing Events in Video Data

Generally, a video sequence consists of ordered frames that can be partitioned into a collection of shots using various image processing techniques like histogram comparisons, motion-based indexing, and optical flow determination. Each shot contains no scene changes and is the basic element for characterizing the video data [14]. Several shots can be grouped logically into episodes or scenes., i.e., an episode is a specific sequence of shots [15]. Several episodes can be put in one sequence. The term *clip* is a generic object without any structural meaning, which is a portion of a video sequence with a starting and ending frame numbers. In order to put things in perspective, we first suggest the following definitions.

Generic indexing : It is the process of identifying a clip from a video sequence and using image processing algorithms (histograms or equivalents) to partition the clip into ordered shots.

Structural indexing : It is the process of grouping continuous shots to form an episode and grouping continuous episodes to form a program.

In this paper we address issues related to structural indexing only. Generally, most of the episodes and programs can be expressed in the form of worldly knowledge by describing the interplay among physical objects in the course of time and their relationship in space. Physical objects may include persons, buildings, vehicles, etc. Video is a typical replica of this worldly environment. In conceptual modeling of video data for the purpose of structural indexing, it is therefore important that we identify physical objects and their relationship in time and space. Subsequently, we can represent these relations in a suitable data structure that is useful for users to manipulate. Temporal relations among objects have been previously modeled by using methods like temporal-interval [9]. For spatial relations, most of the techniques are based on projecting objects on a two or three dimensional coordinate system. Very little attempt has been made to formally express spatio-temporal interactions of objects in a single framework. Though in [8], spatial/temporal metadata for video database is defined, yet no detailed modeling is provided. In the following sections, we describe a generalized framework describing spatio-temporal relationships of objects in an image or video.

2.1 Generalized Spatial and Temporal Operations

Generalized spatial and temporal operations presented in this section are an extension to our earlier work [9]. The reason of introducing the generalization in both spatial and temporal domains is to simplify describing complex spatial or temporal events, which otherwise are rather cumbersome to express [4]. We first give a definition for the generalized n-ary relation.

Definition 1 : Generalized *n*-ary relation *A generalized n-ary relation $R(\tau_1, \ldots, \tau_n)$ is a relation among n objects, τ_is, that satisfies one of the conditions in Table 1 according to their positions in space*

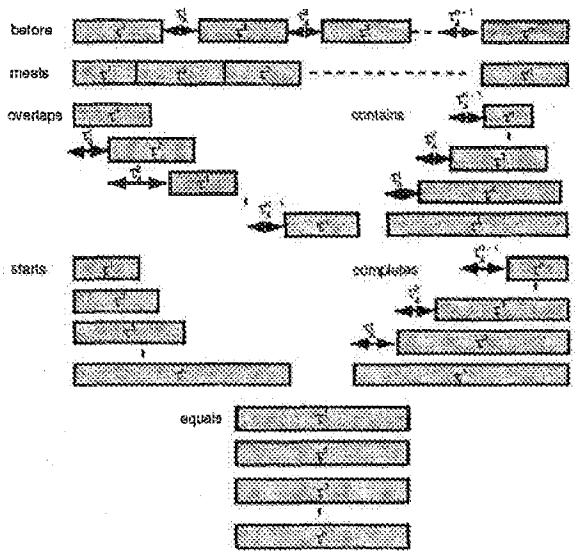

Figure 2: n-ary relations

or time domain with respect to each other.

The relation is represented by the corresponding name and symbol. The operands of the relations, i.e. τ_i, $(i = 1, \ldots, n)$ are either the projections of the positions of the objects (spatial domain) or time span of a certain object/event (temporal domain).

The generalized n-ary relations and the corresponding interval constraints are shown in Figure 2 and Table 1, respectively. The same fundamental relations can be used either in space or time domains. The difference is in the meaning of the operands rather than the operation. For the spatial domain the operands represent the physical location of the objects while in the temporal case they represent the duration of a certain temporal event (such as *presence*). The number of operands, n, in the relations is assumed variable. This generality enables any spatial or temporal situation to be represented in terms of the seven fundamental n-ary relations in Figure 2.

2.2 Modeling of Spatial Events in a Single Frame

Assume that computer vision/image processing algorithms for object identification and recognition have been applied to video frames and a spatial attribute,

Relation name	Symbol	constraints, $\forall\ i,\ 1 \leq i < n$
before	B	$\tau_i{}^e < \tau_{i+1}{}^s$
meets	M	$\tau_i{}^e = \tau_{i+1}{}^s$
overlaps	O	$\tau_i{}^s < \tau_{i+1}{}^s < \tau_i{}^e < \tau_{i+1}{}^e$
contains	C	$\tau_i{}^s < \tau_{i+1}{}^s < \tau_{i+1}{}^e < \tau_i{}^e$
starts	S	$\tau_i{}^s = \tau_{i+1}{}^s \wedge \tau_i{}^e < \tau_{i+1}{}^e$
completes	CO	$\tau_i{}^s < \tau_{i+1}{}^s \wedge \tau_i{}^e = \tau_{i+1}{}^e$
equals	E	$\tau_i{}^s = \tau_{i+1}{}^s \wedge \tau_i{}^e = \tau_{i+1}{}^e$

$\tau_i{}^s$ = starting coordinate of object τ_i
$\tau_i{}^e$ = ending coordinate of object τ_i

Table 1: n-ary relations

called bounding volume, V, for each salient physical object present in the frame has been extracted and stored in VSDG (Video Semantic Directed Graph) [7] or equivalent data structure. The volume describes the spatial projection of an object in x, y, and z axes and is defined in the following way:

$$Bounding\ Volume\ (V) =$$

$$(x_1, x_2, y_1, y_2, z_1, z_2)$$

A 2-D *Bounding Box* is used in those cases where only 2-D information is available. For all three coordinates, the points with subscripts 1 and 2 specify the beginning and end points of the projections respectively.

The information provided by the bounding volumes is not sufficient to describe meaningful semantic information present in a frame. Although it provides the most fundamental information about a frame, e.g., the locations of individual objects, it needs to be expanded to construct higher level contents in the frame. Such detailed information contents in a single frame can be termed as *spatial events*.

For example, *presiding* a meeting attaches a meaning to some spatial area. For this event, a person in a frame may be identified such that he/she is either standing or sitting on a chair in the center or front of a meeting room. Another example of a spatial concept is *three point position* in a basketball field. Similarly, a person may be *sitting* on a chair or some physical object. In this case, we have a conceptual spatial object 'sitting' with attributes 'a physical object which sits' and 'a physical object being sit on', and they are related by the 'sitting' relationship.

In order to express events in an unambiguous way, we present a formal definition of a *spatial event* based on the spatial operations discussed in the previous section.

A spatial event describes the relative positions of objects in a frame.

Definition 2 : Spatial Event *A spatial event E_s is a logical expression consisting of various generalized n-ary spatial operations on projections and is described as follows*
$E_s = R_1(\tau_1{}^1, \ldots, \tau_{n_1}{}^1) \diamond_1 R_2(\tau_1{}^2, \ldots, \tau_{n_2}{}^2) \diamond_2 \ldots \diamond_{m-1} R_m(\tau_1{}^m, \ldots, \tau_{n_m}{}^m)$,
where R_j, $j = 1, \ldots, m$ is a generalized n-ary relation, $\diamond_k$, $k = 1, \ldots, m-1$ is one of the logical operators ($\wedge$ or $\vee$) and τ_j^i is the projection of object j in relation i on x, y, or z axis.

Note that more complex spatial events can be constructed by relating several spatial events using logical operators.

As an example of spatial event, consider a player holding the ball in a basketball game. To simplify the characterization of this situation, we assume when the bounding boxes of the objects player and ball are in contact with each other, it is considered that the frame contains event "player holding the ball". This is characterized by six of the n-ary relations in both x and y coordinates and can be formally expressed as
$E_s = (M(\tau_x{}^1, \tau_x{}^b) \vee O(\tau_x{}^1, \tau_x{}^b) \vee C(\tau_x{}^1, \tau_x{}^b) \vee S(\tau_x{}^1, \tau_x{}^b) \vee CO(\tau_x{}^1, \tau_x{}^b) \vee E(\tau_x{}^1, \tau_x{}^b)) \wedge (M(\tau_y{}^1, \tau_y{}^b) \vee O(\tau_y{}^1, \tau_y{}^b) \vee C(\tau_y{}^1, \tau_y{}^b) \vee S(\tau_y{}^1, \tau_y{}^b) \vee CO(\tau_y{}^1, \tau_y{}^b) \vee E(\tau_y{}^1, \tau_y{}^b))$,
where $\tau_x{}^1$ is the projection of the bounding box associated with object player 1 on the x-axis and $\tau_x{}^b$ is the projection of the bounding box associated with the object ball on the x-axis, etc. If the specified condition is satisfied for a given frame, the event E_s exists.

As a side note, we need to mention that one can maintain the spatial events information for each frame. However, the overhead associated with such detailed specification may be formidable. Also, tracking such detailed information may not even be needed for many applications. We, therefore, can maintain temporal information by only identifying spatial events in frames at δ distance (in frames) apart.

Spatial events can serve as the low level (fine-grain) indexing mechanisms for video data where information contents at the frame-level are generated. Modeling more complex information contents, such as *gloomy weather* is a more challenging problem.

2.3 Temporal Events

The next level of video information modeling involves temporal dimension. Temporal modeling of a video clip is crucial for users to ultimately construct complex views or to describe episodes/events in the clip. Episodes can be expressed by collectively interpreting the behavior of physical objects. The behavior can be described by observing the total duration an object appears in a given video clip and its relative movement over the sequence frames in which it appears. For example, occurrence of a slam-dunk in a sport video clip can be an episode in a user's specified query.

Modeling of this episode requires tracking motion of the player for whom slam-dunk is being queried and tracking motion of the ball in a careful manner especially when it approaches the hoop. Tracking the motion of the player and the motion of the ball are two *simple temporal events*. These temporal events need to be expressed prior to composing the complex episodes of slam-dunk. It is obvious that these simple events can be expressed formally as a temporal sequence of various spatial events, spanning over a number of frames. Composite temporal events are defined in terms of other simple or complex temporal events relating them by the n-ary relations. We formally define a temporal event as follows.

Definition 3 : Temporal Events *A simple temporal event (E_{st}) is defined as a logical operation on a set of spatial events the durations of which are related by one of the n-ary temporal relations. Formally,*
$E_{st} =$
$R_1(d(E_{s_1}), \ldots, d(E_{s_n})) \diamond_1 R_2(d(E_{s_1}), \ldots, d(E_{s_n})) \diamond_2 \ldots \diamond_{m-1} R_m(d(E_{s_1}), \ldots, d(E_{s_n}))$,
where R_j is a generalized n-ary relation and $d(E_{s_i})$ is the duration of the spatial event E_{s_i}. A composite temporal event (E_{ct}) is formed by further relating the existing temporal events using the same spatio-temporal generalized operators. Formally,
$E_{ct} =$
$R_1(d(E_{t_1}), \ldots, d(E_{t_n})) \diamond_1 R_2(d(E_{t_1}), \ldots, d(E_{t_n})) \diamond_2 \ldots \diamond_{m-1} R_m(d(E_{t_1}), \ldots, d(E_{t_n}))$,

where $d(E_{t_i})$'s in this case are durations of temporal events which could be either simple or composite.

In video data, associated with each spatial event is its duration $d(E_s)$ during which the spatial event persists. If the event starts at frame # α and ends at frame # β then $d(E_s) = \beta - \alpha + 1$. The duration of the result of an n-ary operation is the aggregate duration, i.e. the time interval between the earliest starting time and the latest ending time of the involved objects.

A set of spatial/temporal events can be arranged in a nondecreasing order in terms of the (approximate) start time. However, we may not know the exact inter-interval delays in many cases during the definition of

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.