31-1-2000

Semantic modeling for video content-based retrieval systems

L. A. Al Safadi
University of Wollongong, uow@alsafadi.edu.au

J. R. Getta
University of Wollongong, jrg@uow.edu.au

Semantic modeling for video content-based retrieval systems

Abstract

This work proposes a semantic data model for video documents based on the story-line structure powerful enough to express various human interpretations of video documents, and introduces a formal query language for video retrieval that facilitates retrieval of users' heterogeneous queries based on the proposed model. The paper identifies the elementary semantic units, composite semantic units, associations and abstraction mechanisms necessary for symbolic modeling of semantic video contents. The method is independent of presentation media and it has its origins in symbolic modeling systems developed for database and complex software systems design.

Keywords

content-based retrieval, data models, query languages, video databases

Disciplines

Physical Sciences and Mathematics

Publication Details

This paper originally appeared as: Al Safadi, LA and Getta, JR, Semantic modeling for video content-based retrieval systems, 23rd Australasian Computer Science Conference, 31 January - 3 February 2000, 2-9.

Semantic Modeling for Video Content-Based Retrieval Systems

Lilac A. E. Al Safadi and Janusz R. Getta
School of Information and Computer Science
University of Wollongong, Wollongong, NSW 2522
Australia
{laa03,jrg}@cs.uow.edu.au

Abstract

This work proposes a semantic data model for video documents based on the story-line structure powerful enough to express various human interpretations of video documents, and introduces a formal query language for video retrieval that facilitate retrieval of users' heterogeneous queries based on the proposed model. The paper identifies the elementary semantic units, composite semantic units, associations and abstraction mechanisms necessary for symbolic modeling of semantic video contents. The method is independent on presentation media and it has its origins in symbolic modeling systems developed for database and complex software systems design.

1. Introduction

The advances of multimedia technologies enable electronic processing of information recorded in the formats different from a standard text format. These include image, audio and video formats. Video format is a rich and expressive media used in many areas of our everyday life like education, medicine, engineering, etc. Expressiveness of video documents is the main reason of their domination in future information systems. Efficient access to large video stores needs more sophisticated *content-based* video indexing and retrieval systems to retrieve a set of video clips from a large collection on the basis of the content description. While a great deal of effort has been invested into general video data retrieval, relatively little has been done in the area of *semantic content-based* video retrieval which aim to convey elements of meaning that are beyond image, voice, and video analysis, and describe video documents in terms of real world objects, their properties, relationships between them, actions performed by objects, synchronisation of their actions, classification, aggregation and generalization abstractions.

Current trends in video analysis systems use a technique of video stream segmentation in a way either depending on its physical structure (frames or pixels) or its screenplay structure like for instance extraction of representative key frames, identification of scenes or episodes, semantic classification of segments, etc. [10], [13], [21], [22]. These approaches are not closely related to video semantics and because of that they do not capture the underlying semantic structures based on user's view. Existing semantic models are too poor for video modeling. [18, 19] Are limited to keywords match, and do not address relationships between contained semantics. We argue that associations are important in modeling real world, and that these associations build other high level semantic units. [3, 8] Develop a video data model consisting of frame-based objects and relationships. As present, database world is into a high level descriptions of multimedia. Semantics, concurrently, must account on how primitive semantics are combined to form the meaning of the whole video stream. Our approach uses an indexing technique based on reconstruction of semantic level through video story-lines built from objects, activities, events, and associations. The experiments conducted so far proved that video indexing systems solely based on signal and image processing are inadequate to a task of representing the semantic contents of video documents [7], and that keyword-based indexes for large video stores can be easily implemented and effectively used [2],[5], [13], [24], [25]. Several techniques has been proposed for video indexing such as scene cut detection [16] to detect semantic boundaries, automatic extraction of visual features and capturing static salient object [9], objects in motion [5], identify faces appeared [23], and embedded captioned information [12]. Audio processing techniques managed capturing and classifying non-speech sounds [24], as well as spoken words [4].

With these techniques, it is possible to: 1) Identify the groups of pixels as selected predefined objects like person, car, etc. 2) Identify a dynamic object's movement like run, walk, etc. 3) Address particular properties existed in the perceptual level like colour, texture, etc. and 4) Address explicit spatial and temporal relationship. However, these techniques are not sufficient for recognizing: 1) Content properties beyond the perceptual level like name, age, etc. 2) Implicit association which not explicitly listed but required to form an abstract unit, and 3) Complex semantic units like events and stories told by video documents. Our work suggest a human-centred video analysis system based on the approach so successfully exercised by many Computer Aided Design (CAD) systems. In a CAD system a human operator plays a central role in the design process while a machine is used when the complex and well-defined computations should be done.

In a human-centred video indexing system, image and audio processing techniques are only applied to extract perceptual features and capture physical level semantic units (objects and primitive activities) and relationships (spatial and temporal attributes). Then, a human operator uses his knowledge and experience to identify the elements that is hard to discover from pixels and frequencies, i.e. implicit relationship between semantic units, complex semantic units, and their behaviour. The aim of this work is to set a semantic structure, which is an essential part toward automating video semantic contents.

So far simple systems such as the one proposed by Courtney in [5], makes a small step towards machine supported human analysis of video documents.

This paper proposes a semantic modeling of video documents. It is based on a story-line structure. The model is powerful enough to express various human interpretations of video documents, and introduces a query language for video retrieval. Moreover, the language is able to process heterogeneous queries based on the proposed model.

Our paper is organised as follows. Section 2 contains a brief analysis of the basic concepts of our model. The proposed model is defined in section 3. Section 4 presents a formal query language. Section 5 introduces a graphical user interface for video contents. Finally future work and summary in section 6.

2. The semantic modeling of video documents

A logical view of a typical video document is a recorded sequence of frames (images). When the frames are displayed fast enough, the small discrepancies in the positions of displayed objects create an illusion of their movements. A semantic view of a video document is a recorded sequence of events that have happened in the real (sometimes imaginary) world.

Semantic modeling of video documents requires a formal symbolic modeling system typically called by database designers as conceptual or semantic model. Symbolic modeling has been used in many areas like for instance conceptual modeling for database systems, knowledge acquisition and representation in artificial intelligence, modeling of complex software systems in software engineering. The same video document may have different user's view. A *user view* is the perception of what is the contents of the proposed video. The methodology for generating user views is that, video expose a number of semantic units. People would reveal different perspective depending on units and descriptions of their interest, association, and their level of abstraction. Hence, our conceptual model constitutes:

(i) semantic units,

(ii) descriptions of semantic units,

(iii) associations between semantic units,

(iiii) abstraction mechanisms over semantic units.

The choice of semantic units determines the expressiveness, completeness, and flexibility of semantic model. Our comprehension of video documents as recorded sequences of events involving real world objects and actions performed by them indicates the choice of objects and actions as elementary semantic units. An obvious conclusion that objects perform actions and they are related to each other leads to a concept of composite semantic units. A *composite semantic unit* is a structure built of the elementary and possibly composite semantic units to express the complex facts, for instance an action is associated with another object like drinking beverage, a group of objects collaboratively performs an action like team working, etc. We introduce the following levels of granularity of composite semantic units:

(i) events which consist of objects, actions, and associations observed in a period of time,

(ii) stories which are made of a number of events.

(iii) complex semantic unit which consist of other elementary or composite units observed within a period of time.

An important feature of a semantic model is its ability to describe the properties of semantic units. In our system a description of semantic unit consists of a name, and set of single and/or multi-valued attributes. For example an event named as SIGMOD99 has the attributes start/end dates, location and belongs to conference class. Attributes could be dynamic or static. *Dynamic* attribute changes its value over time. *Static* attribute has a fixed value.

We assume that every semantic unit in a video passes through successive states. A transition from state to state is determined by a change in the value of an observable dynamic attribute. Each state has an associated video identifier VID, and pair of two numbers (frame numbers or time) to represent a period of time when a unit is observable (from a moment t_s to a moment t_e). A triple $[VID, t_s, t_e]$ is called an *observation slot* of semantic unit state. Observations slots of all components should be included within observation slot of a unit they belong to. A concept of observation slot links an abstract concept of semantic unit with a physical chunk of video document.

Abstraction is a mental process that leads from identification of instances of semantic units and their descriptions to identification of the homogeneous groups of semantic units later on called as classes of semantic units. Three typical abstractions include classification, generalization, and aggregation. *Classification* abstraction is used to discover the classes of semantic units from information about instances of semantic units, e.g. the instances of car drivers form a class driver, the instances of events like SIGMOD98, VLDB97, IMC99 form a class of events called conference. Classification abstraction may be applied to any type of semantic unit i.e. objects, actions, events, and stories. Each class of semantic units has a description that consists of name and set of attributes. *Generalization* abstraction leads to identification of inclusion relationship among homogeneous classes of semantic units. For example we discover that class of objects driver is a subclass of class person and class of actions to-run is a subclass of class to-move. *Aggregation* abstraction is a structuring mechanism for assembling complex semantic units from the elementary ones. For instance, in our model aggregation abstraction is used to construct associations from objects and actions, events and stories. It may also be used to construct complex objects from elementary objects or complex actions from elementary actions. It seems that complete description of a single even quite short video document by a human viewer is impossible because of a huge number of indexing aspects that may be addressed by a viewer.

3. The model

This section provides a formal specification of elementary semantic units, associations, their descriptions and abstractions over semantic units.

3.1. Formal definition of semantic unit

A semantic unit is a quadruple (uid F, V, ∂), where uid is the semantic unit identifier, F is a set of content attributes, and V is a set of attributes' values $V = \bigcup_{f \in F} domain(f)$. Then ∂ maps attributes into their values $\partial : F \longrightarrow V$ such that $\partial(f) \in domain(f)$
For instance, suppose we have an object person with a quadruple (123, F, V, ∂) where:
F = { name, date-of-birth, shirtcolor, class, ... } is a set of content attributes.
V = { Ali, 2-5-1970, red, person, ... } is a set of attributes' values.
∂(name) = Ali, ∂(dat-of-birth) = 2-5-1970

3.2. Objects

An instance of physical object is any salient object captured in a video's physical space represented visually, aurally, or textually. A physical object becomes a semantic object when a viewer identifies a class of real world objects a physical object belongs to. A video may contain a number of salient objects not recognised by a viewer, hence they remain only physical objects and not semantic. Throughout our work, we will use the term *object* to refer to an instance of semantic object. Every object in the indexing system obtains a unique name and optional attributes.

3.3. Actions

Observation of continuing changes in the values of object's dynamic attributes over an interval of time is interpreted as an *action*. A semantic action is an identified action performed by an object which could be referred to as the *actor* performing the action. Action and actor are associated in a 1:1 performed-by relationship, and denoted by $A(O)$ where A is an action, and O is the actor. Actions are described in the same way as objects, where they have unique name and optional attributes.

3.4. Associations

Semantic units in a video are related in an n-ary semantic space. For instance in "man drink

beverage", there is a concealed relationship between the action (drink) performed by actor (man) and object (beverage), or explicit connection, for instance "X *father-of* Y". Semantic association is denoted by $R_X(A_1, \ldots, A_n)$ where A is a semantic unit and $R_X \in \{father-of, friend-of, \ldots\}$. Two set of classes maybe distinguished in semantic association, spatial and temporal associations. *Spatial* association is a binary association between two semantic units indicating relationship in space, and denoted by $R_S(A_1, A_2)$, where $R_S \in \{above, left, infront, between\}$. For instance, "book above table" is a spatial association between two objects. "Accident behind the bridge" is a spatial association between an event and an object. *Temporal* association is a binary association between two semantic units interpreted in time, and denoted by $R_T(A_1, A_2)$, where $R_T \in \{$ *before*, *meet*, *during*, *overlap*, *starts*, *ends*, *equal*$\}$. Our choice of temporal associations comes from [1] . For instance, "man runs after a dog" is a temporal relationship between an action and an object in time. Associations has unique name, and optional attributes.

3.5. Events and stories

Event is an interpretation of a number of contextually related activities, objects, and associations, denoted by $E(A, S)$, where A a set of actions or objects, and S is a set of associations such that $\forall a_i \in A, \exists a_j \in A$ and $\exists s \in S$ where $s(a_i, a_j)$ and i $\neq$ j. The function F maps objects, actions, and associations into an events, $F : (A, S) \longrightarrow E$.
Example consider a sequence of frames representing a leaving event given in Figure 1. Changes in the spatial parameters of two objects over a sequence of frames are captured. There are two objects o_1 of class person and o_2 of class door and three actions: a_1 of class to-walk, a_2 of class to-open performed by o_1 and implicitly associated to o_2, and a_3 of class swing performed by o_2. The term *event* in our work refers to semantic event. It represents an abstraction of a collection of bounded objects and activities. Some conceptual models define event is an instant of occurrence while the others define event as what triggers an action. Most works do not differentiate between actions and events. Story is a sequence of events $[e_1, e_2, \ldots, e_n]$.

3.6. Abstractions

Three common abstraction mechanisms classification, generalization and aggregation abstractions

Figure 1: A sequence of frames representing a "leaving" event.

are available for grouping instances of semantic units instances within classes, building class hierarchies and construction of complex semantic units. Classification abstraction allows for defining the classes of semantic units e.g. class of objects person, class of actions running, class of events conference, etc. Let C be a set of homogeneous classes of semantic units, e.g. a set of all object classes. Then, generalization abstraction G is defined as subset of $C \times C$. Generalization abstraction allows for defining the hierarchies of semantic units classes like for instance postgraduate-student class is a subset of student class which is on the other side a subset of person class etc. Aggregation abstraction can be used to define the complex classes of semantic units. For instance an object of class car is an aggregation of more elementary objects from the classes like wheel, engine, chassis, etc.

3.7. Formal definition of a semantic unit and association in video

The formal definition of semantic unit presented earlier in section 3.1. Semantic units and association in a video are recorded in a 7-tuple (S, *uid*, T, F, V, φ, λ), where S is a set of state identifiers, *uid* is the semantic unit or association identifier, T is a set of observation slots triple $[vid, t_s, t_e]$, F is a set of a dynamic attributes, V is the set of their values, φ maps states into set of attributes and values such that $\varphi : S \longrightarrow P(\partial)$ and $\varphi(s) \in \{ \partial_1, \partial_2, \ldots \}$ where $\partial \in \partial$, and λ maps states into observation slots such that $\lambda : S \longrightarrow T$ then $\lambda(s) \in t$
For instance, suppose we have the a semantic object person with a 7-tuple (S, 123, T, F, V, φ, λ) where:
$S = \{ s_1, s_2, \ldots \}$ set of unit's states.
$T = \{$ [222, 20, 45], [222, 70, 95], [333, 120, 127], $\ldots \}$ set of observation slots where object appeared in.
$F = \{$ shirtcolor, X, Y, $\ldots \}$ are set of dynamic attributes.
$V = \{$ red, white, 20, 30, 40, 45, $\ldots \}$ are set of attributes' values.

∂_1(shirtcolor)= red, ∂_2(X) = 30, ∂_3(shirtcolor) = white, ∂_4(Y) = 45
$\varphi(S)$ maps states into attributes and attributes' values as follows:
$\varphi(\ s_1\ ) = \{\ \partial_1, \partial_2\ \}$, $\varphi(\ s_2\ ) = \{\ \partial_3, \partial_4\ \}$
$\lambda(S)$ maps states into observation slots as follows:
$\lambda(s_1)$ = [222, 20, 45], $\lambda(s_2)$ = [222, 70, 95]

4. Query language

In this section we present a formal query language based on the first ordered logic notation to build queries to video database [15]. Our idea is to build query language using $\neg(not), \wedge(and), \vee(or), \forall(for\ all), \exists(there\ exist), |$ (*such that*), set of predicates, functions, constants (e.g. 123, red, Ali, ...), and variables representing semantic units and attribute values (e.g. x, y, ...). Parentheses are used to override the precedence of the symbols. A number of predicates and functions are defined in our system: class, association, description, and semantic structure predicates. Some association functions and description predicates maybe created automatically from existed identified semantic units and associations. In other word, the identification of new class, association, or description, automatically has its impact on query language by obtaining a new predicate or function.

1. *Class* predicate written in upper-case letters, identifies the class in which a semantic unit belongs to. For instance, STUDENT(x). For generalization abstraction, concepts are organized into a hierarchy of IS-A relationship, where subclasses inherit all properties of superclasses. For instance, an postgraduate student IS-A student, and an student IS-A person.. etc. the hierarchy leafs corresponds to specific concepts (postgraduate), and higher nodes corresponds to more fuzzy unspecified concepts (person). Fuzzy concepts are relaxed by adding to the query posed to video database the set of possible concepts that a unit could represent. For instance, a student is relaxed into postgraduate, undergraduate. STUDENT(x) $\Rightarrow$ (POSTGRADUATE(x) $\wedge$ UNDERGRADUATE(x))

2. Association functions driven automatically from registered semantic units and associations. Some *temporal* function (before, meet, during, overlap, starts, ends, equal) , and *spatial* functions (below, left, in front, between) are predefined. Concealed semantic association is identified by *ass* predicate. Association functions are denoted by $f(u_1, u_n)$, where n=2 in temporal and spatial associations.

3. *Description* predicates associate semantic unit with attribute values representing attribute name. For instance, color(x, red). Predicates indicates equality unless elsewhere specified. For instance, age(y, greater, 20).

4. *Semantic structure* predicates actor(x, y) and comp(x, y). actor(x, y) returns TRUE if the dynamic object x performs the action y. comp(x, y) boolean predicate returns TRUE if x is a sub-component of a composite unit y (not important a direct component). For instance, lecturing action is a component-of speech events, and that is a component-of conference. However, comp(lecturing, speech), and comp(lecturing, conference) , all return TRUE. Both semantic structure predicares implies partial observation slot ordering. In other word, actor(x, y) and comp(x, y) implies $x.t \subseteq y.t$ where t is the observation slot.

4.1. Query language examples

Example 1. Red cars is expressed as

$$\{\ x \mid \text{CAR}(x) \wedge \text{color}(x, \text{red})\ \}$$

This example retrieves a semantic object belong to class CAR, and described by having a red color attribute.

Eaxmple 2. A man walking and not car running

$$\{\ x \mid \exists\ y\ j\ k\ (\ \text{MAN}(j) \wedge \text{WALK}(x) \wedge \text{actor}(j, x) \wedge \text{CAR}(k) \wedge \text{RUN}(y) \wedge \text{actor}(k, y) \wedge \text{overlap}(x, \neg\ y)\ \}$$

Query criteria in this example aim to retrieve a video clip where the two actions, walk performed-by a man and run performed-by a car, do not appear simultaneously.

Example 3. Conference where editorial presented by Ali followed by a Multimedia lecture

$$\{\ x \mid \exists\ y, j, k, z \Big(\ \text{CONFERENCE}(x) \wedge \text{comp}(y, x) \wedge \text{comp}(z, x) \wedge \text{comp}(\text{before}(y, z), x) \wedge \Big(\ \text{EDITORIAL}(y) \wedge \text{comp}(j, y) \wedge \big(\ \text{PRESENT}(j) \wedge \text{actor}(k, j) \wedge (\ \text{PERSON}(k) \wedge \text{name}(k, \text{Ali})\ )\ \big)\Big) \wedge \Big(\ \text{LECTURE}(z) \wedge \text{subject}(z, \text{Multimedia})\ \Big)\ \Big)\ \}$$

This is a query of a composite event (CONFERENCE) composed of EDITORIAL and LECTURE subevent.

Table 1: Association Predicates Interpretation

Predicate	Interpretation
A before B	$A.t_e < B.t_s$
A meet B	$A.t_e = B.t_s$
A during B	$A.t_s \geq B.t_s$ and $A.t_e \leq B.t_e$
A overlap B	$A.t_s \leq B.t_s$ and $A.t_e \leq B.t_e$ or B overlap A
A starts B	$A.t_s = B.t_s$
A ends B	$A.t_e = B.t_e$
A equal B	$A.t_s = B.t_s$ and $A.t_e = B.t_e$
A left B	$A.x < B.x$
A above B	$A.y > B.y$
A in front B	$A.z < B.z$
A between B	$A.x \geq B.x$ and $A.x+width \leq B.x+width$ and $A.y \geq B.y$ and $A.y+height \leq B.y+height$

The EDITORIAL event is constituted of PRESENT action performed by an object of class PERSON.

4.2. Query interpretation

A query Q submitted to video content-based retrieval system, is a composition of n associated variables $< x_1, \ldots, x_n >$ resulting n tuples $\{< r_1, \ldots, r_n > : f(r_1, \ldots, r_n)$=TRUE$\}$. Each tuple r is a result of a number of matched content criterias $(f_1 : domain(f_1) \wedge \ldots \wedge f_n : domain(f_n))$. Table 1 list the interpretation of spatial and temporal association functions based on semantic unit's spatial and temporal attributes. Follows is the evaluation of queries listed in section 4.1 above.

Example 1. Red cars query returns the tuple:
(s, uid, t, { class, color }, { car, red }, φ, λ) where:
∂_1(class) = car, ∂_2(color) = red
φ(s) = { ∂_1, ∂_2 }, $\lambda(s) = t$

Eaxmple 2. A man walking and not car running
(s_1, uid, t_1, {class}, {man}, φ, λ)
∂_1(class) = man
φ(s_1) = { ∂_1 }, λ(s_1)=t_1

(s_2, uid, t_2, {class, actor}, {walk, s_1}, φ, λ)
∂_1(class) = walk, ∂_2(actor) = s_1
φ(s_2) = { ∂_1, ∂_2 }, λ(s_2) = t_2

(s_3, uid, t_3, {class}, {car}, φ, λ)
∂_1(class) = car
φ(s_3) = { ∂_1 }, λ(s_3) = t_3

(s_4, uid, t_4, {class, actor}, {run, s_3}, φ, λ)
∂_1(class) = run, ∂_2(actor) = s_3
φ(s_4) = { ∂_1, ∂_2 }, λ(s_4) = t_4

(a, uid, t, {name, operand1, operand2}, {overlap, s_2, s_3}, φ, λ)
∂_1(name) = overlap, ∂_2(operand1) = s_2, ∂_3(operand2) = s_3
φ(a) = { ∂_1, ∂_2, ∂_3 }, λ(a) = t

Example 3. Conference with ediorial presented by Ali followed by a Multimedia lecture query is decomposed into a number of subqueries.
Based on the query structure, the final query is composed as follows:
F : (EDITORIAL, LECTURE, before) $\longrightarrow$ CONFERENCE
F : (PRESENT(PERSON)) $\longrightarrow$ EDITORIAL
We start by evaluating the query in a bottom-top manner.
(s_1, uid, t_1, {class, name}, {person, Ali}, φ, λ)
∂_1(class) = person, ∂_2(name) = Ali
$\varphi(s_1)$ = { ∂_1, ∂_2 }, $\lambda(s_1) = t_1$

(s_2, uid, t_2,{class, actor},{present, s_1}, φ,λ)
∂_1(class) = present, ∂_2(actor) = s_1
$\varphi(s_2)$ = { ∂_1, ∂_2 }, $\lambda(s_2) = t_2$

(s_3, uid, t_3,{class, comp},{editorial, s_2},φ,λ)
∂_1(class) = editorial, ∂_2(comp) = s_2
$\varphi(s_3)$ = { ∂_1, ∂_2 }, $\lambda(s_3) = t_3$

(s_4, uid, t_4, {class, subject}, {lecture, Multimedia}, φ, λ)
∂_1(class) = lecture, ∂_2(subject) = Multimedia
$\varphi(s_4)$ = { ∂_1, ∂_2 }, $\lambda(s_4) = t_4$

(a, uid, t, {name, operand1, operand2}, {before, s_3, s_4}, φ, λ)
∂_1(name) = editorial, ∂_2(operand1) = s_3, ∂_3(operand2) = s_4
$\varphi(a)$ = { ∂_1, ∂_2, ∂_3 }, $\lambda(a) = t$

(s_5, uid, t_5, {class, comp1, comp2, comp3}, {conference, s_3, s_4, a}, φ, λ)
∂_1(class) = conference, ∂_2(comp1) = s_3, ∂_3(comp2) = s_4, ∂_4(comp3) = a
$\varphi(s_5)$ = { ∂_1, ∂_2, ∂_3, ∂_4 }, $\lambda(s_5) = t_5$

5. Graphical conceptual model for video contents

The aim of this section is to introduce graphical model to describe the interplay among semantic units constituting a composite unit, which we believe it can be a step toward an easy to grasp graphical user interface where for each input video stream, semantic units are captured and encoded on a proposed graphical model. Object Composition Petri Net OCPN model [13], which we decided to adopt in our work, is suitable for representing concurrence and synchronisation. In OCPN, circles represent a state of a component. State modification is associated with the change in presentation time. Duration is assigned to each state representing the time interval in which a state is active. Vertical bar represents a transition or point of synchronisation, describing when do components synchronise their presentation, and projecting the temporal order of components.
An event is a composite semantic unit of one or more synchronised *related* components, which impose synchronisation and relationships in presentation. One of the limitations of OCPN is that it does not express all semantic relationships between components. Only temporal relationships. On the other hand, Entity-Relationship model [20], is a very efficient graphical conceptual model in representing the relationship between entities, but fails to express synchronisation. Therefore, the formal definition of our proposed model is that, it is a direct graph, adopted from OCPN, and extended by adding a temporal bi-directional lightning arrow to describe a relationship between two components. This relationship symbol, requires the presence of both component. In other words, the removal of one or a change in state will lead to the termination of relationship. To illustrate our idea, consider the sequence of frames representing a leaving event described in Figure 1. The graphical representation of the event is given below (Figure 2). Object o1 appears at moment t1 performing action a_1, object o_2 appears at t_2, at t_3 o_2 is involved in association with action a_2 performed by o_1 $(a_2(o_1) : o_2)$, and action a_3 performed by o_2 denoted by $a_3(o_2)$ appears at t_4, then o_1 disappears at t_5.

6. Summary and future work

Representation of semantic contents of video documents is needed for construction of the large robust multimedia stores, which support content-based manipulation of video, audio and images. The approaches solely based on signal scanning and parsing

Figure 2: Graphical representation of a "leaving" event.

are inadequate to this task, and do not address the underlying semantic structure of a video. We believe with today's automatic analysers, it is impossible to capture high level video semantics. Therefore, we propose a human-centred architecture of content-based video indexing systems where a human operator supported by processing software systems plays a central role in the semantic indexing of video documents. One of the first steps towards implementation of this idea is a formal specification of conceptual model. The objective of this work was to define a conceptual model powerful enough to describe the semantic contents of video documents that greatly facilitate heterogeneous queries. The model is based on three concepts: elementary semantic units, descriptions, associations, and abstractions. Descriptions register the states of instances of semantic units, and classes of semantic units. Abstractions enable classification of semantic units, reasoning about classes of semantic units and construction of complex semantic units. Our approach extends a plethora of already proposed symbolic modeling tools by recognition of elementary concept of action and in consequence by allowing associations (relationships) to be defined over both objects and actions. Another extension allows for application of abstraction mechanisms to any type of semantic units and not like in the other models only to objects. Synchronisation mechanisms needed for dynamic description of related actions are provided by representation of events as Petri nets. The model identifies a concept of description as one of its three main components. Autonomy of description allows for its application not only in the context of objects and associations, but also to any semantic unit as well as to classes of semantic units and logical streams.

We are aware that our open semantic content-based annotation has some constraints that should be considered while retrieval:

1. Semantic units in our work are schemaless where any content attribute can be defined, hence, an important task is to identify user's predicate which internally represents an attribute name

that is invisible to user.

2. Semantic information has some constraints due to its domain-dependencies, synonyms, homonyms, various level of abstraction, and users' query heterogeneity.

3. Due to semantic information constraints, incomplete video annotation, real-world fuzzy information, and imprecise user's query, video retrieval should reason with fuzziness.

Therefore, we need to develop an algorithm for matching query with video extracted contents that supports open set of predicates, fuzzy matching, query relaxation, and user's domain identification. Our future work aims to map proposed data model into relational database, present and implement video retrieval algorithms.

References

[1] Allen, J. F.: Maintaining knowledge about temporal intervals. Communications of the ACM (1983) 832–843

[2] Adali, S., Candan K. S., Chen, S., Erol, D. K., Subrahmanian, V. S.: The Advanced Video Information System. Multimedia Systems **4** (1996) 172-186

[3] Aghbari, Z., Kaneko, K., Makinouchi, A.: VST-Model A Unified Topological Modeling of the Visual-Spatio-Temporal Video Features, Proceedings of IEEE International Conference on Multimedia Computing and Systems (1999)

[4] Brown, M. G., Foote, J. T., Jones, G. J., Jones, K. S., Young, S. J.: Open Vocabulary Speech Indexing for video and Audio Mail retrieval. Proceedings of the Fourth International ACM Multimedia Conference **96** (1996) 307–315

[5] Courtney,J. D.: Automatic, Object-Based Indexing for Assisted Analysis of Video Data. Proceedings of the Fourth International ACM Multimedia Conference **96** (1996) 423–424

[6] Dimitrova, N.: The Myth of Semantic Video Retrieval. ACM Computing Surveys, **27** (1995) 584–586

[7] Davis, M., Baudin, C., Kedar, S., Russell, P.M.: No-Multimedia without Representation. Proc. of the Second ACM Intl. Conf. On Multimedia (1994) 181–183

[8] Decleir, C., Hacid, M.: Modeling and Querying Video Data A Hybrid Approach, In Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries (1998)

[9] Flickner, M.,Sawhney, H., Noblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yankar, P.: Query by Image and Video Content: The QBIC System. IEEE Computer, September (1995)

[10] Hjelsvold, R., Midtstraum, R.: Modeling and Querying Video Data. Proceedings of the VLDB Conference (1994)

[11] Kaindl, H., Carroll, J.M.: Symbolic Modeling in Practice. Communications of the ACM, **42** (1999) 28–30

[12] Lienhart, R., Automatic Text Recognition for video Indexing. Proceedings of the Fourth International ACM Multimedia Conference **96** (1996) 11–20

[13] Li, J.Z., Goralwalla, I.A., Özsu, M.T., Szafron, D.: Video Modeling and its Integration in a Temporal Object Model. TR 96-02 The University of Alberta (1996)

[14] Little, T. D. C., Ghafoor, A.: Synchronization and Storage Models for Multimedia Objects. IEEE Journal on Selected Areas in Comm., **8** (1990) 413–427

[15] Maier, D.: The Theory of Relational Databases. Pitman Publishing Ltd. (1983)

[16] Meng, J., Juan, y., chang, S.: Scene Change Detection in a MPEG compressed video sequence. IS&T/SPIE Symposium Proceedings, Vol. 2419 (1995)

[17] Nakamura, Y., Kanade, T.: Semantic Analysis for Video Contents Extraction, Spotting by Association in News Video. Proc. of the Conference on Multimedia (1997) 393-401

[18] Oomoto, E. , Tanaka, K.: OVID: Design and Implementation of a Video-Object Database System , In Proceedings of IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 4, Aug. 1993, P. 629 - 643

[19] Subrahmanian, V. S.: Video Databases, Principles of Multimedia Database Systems (1998)

[20] Storey, V., Goldstein, R.: A Methodology for creating User View in Database Design. ACM Transactions on database systems, Col. 13, No. 3 (1988) 305–338

[21] Swanberg, D., Shu, C., Jain, R.: Knowledge Guided Parsing in Video Database. Electronic Imaging: Science and Technology, San Jose, California (1993)

[22] Sistla, A. P., Yu, C., Venkatasubrahmanian, R.: Similarity Based Retrieval of Videos. IEEE Proceedings 13th International Conference on data Engineering (1997)

[23] Wu, J., Aug, Y., Lam, P., Moorthy, S., Narasimhalu, A.:Facial Image Retrieval, Identification and Inference system. Proc. of ACM Multimedia (1993) 47–55

[24] Wold, E., Blum, T., Keislar, D., Wheaten, J.: Content-Based Classification, Search, Retrieval of Audio. IEEE Multimedia, **3** (1996) 27–36

[25] Ya, H., Wolf, W.: A Visual Search System for Video and Image Databases. Proceedings of the IEEE International Conference on Multimedia Computing and Systems (1997) 517–524