Video Semantic Models : Survey and Evaluation*

Yu Wang, ChunXiao Xing and Lizhu Zhou

Tsinghua University, Beijing, China

Summary
With the development of video technology and appearance of new video-related applications, the amount of video data has increased dramatically which demands support in semantic models to facilitate information representation and query. The video semantic models surveyed in this paper are classified into two categories: annotation-based models and rich semantic models. However, currently there are no criteria for a good semantic model so people lack the rules for evaluating an existing model and the guidelines for designing a new model when necessary. To address this issue, this paper proposes twenty one properties as the criteria for video semantic models, and evaluates eleven existing rich semantic models according to these properties. The results show that these models mostly concentrate on aspects of basic expressive power and query ability. But for some advanced features such as user-defined constraints, assistance for acquisition of semantic information, query evolution etc., there are rooms for further enhancement. The paper concludes by analyzing the evaluation results and indicating research directions for future video semantic models.

Key words: *Video semantic model, evaluation*

1. Introduction

Development in video technology and occurrence of new video applications has lead to an enormous growth in the volume of video data. However, these data are useful for end users only when they can find what they need efficiently and accurately. To achieve that, an appropriate data model is needed since traditional data models cannot fulfill the requirements for video data featured by complex structure and rich semantics.

Information conveyed by video data may be classified into three categories. Low-level feature information includes features such as color, texture, shape etc. Syntactic information describes **what is in the video**, including salient objects, their spatio-temporal position and spatil-temporal relations between them. Semantic Information describes **what is happening in the video** and is perceived by human users.

To illustrate the difference of the three kinds of information, let's consider a video segment in a NBA game in which YaoMing passes the ball to Sura. The low-level feature information includes color, texture, motion, etc. The syntactic information includes three salient objects (two of them correspond to the two players (denoted as A and B) and the third corresponds to the ball (denoted as C)) and the moving of C from A to B. The semantic information may be an event named "Passing" with two roles being Yaoming and Sura.

For each kind of information, a data model is needed to manage them efficiently and provide necessary query ability. Many models have been proposed considering each of the three kinds of information. The low-level feature models[1,2,4] use automatically extracted features to represent the video. Syntactic models (spatio-temporal models[5]) represent video content by spatio-temporal positions and relationships, and moving trajectory, which is very limited and awkward for end users. Semantic models[6-29] represent information perceived by human users when understanding the content of video. This information is human oriented and cannot be extracted automatically.

Queries are issued by designating conditions about the three kinds of information. Apparently, query by semantics is the most nature and straightforward way for end users. Thus, in this paper we will concentrate on semantic models of video.

Since video semantic model was introduced in early 90's, it has developed from annotation-based model to rich semantic model. The annotation-based models use text annotated video data to represent video semantics. Its expressive power and query capability are both limited. In contrast to annotation based model, rich semantic model is much powerful in these two aspects. They can represent the real world objects appearing in the video, the abstract concepts such as events, or even those that do not appear but are implied in the video such as background knowledge. Queries like "give all the shots in NBA games that a team performed a quick attack lasting less than 5

* Supported by the National Natural Science Foundation of China under Grant No.60473078

Manuscript received August, 2005.
Manuscript reviced September, 2005.

seconds" can be answered by the query language of such models.

Although various rich semantic models have been proposed, however, no evaluation criteria have been presented to judge the power of a model. This observation motivates the writing of this paper. In this regard, the paper makes main contributions by:

- A survey on typical existing video semantic models and classification of these models into two categories: annotation-based models and rich semantic models;
- A proposal of twenty one evaluation rules for rich semantic models covering expressive power, query capability, and supporting for facilitating acquirement of semantics. Evaluation of existing rich semantic models according to these rules; and
- Analysis of the evaluation results, presentation of developing trends of video semantic models, and indication of future research directions.

This paper is structured as follows. Section 2 is a review of existing video semantic models. Section 3 gives a set of evaluation rules and the result of application of these rules to some typical existing models. Section 4 analyzes the evaluation results. Section 5 is the conclusion.

2. Survey of Existing Models

In this section we will give a survey of existing video semantic models, which are divided into two categories: annotation-based model and rich semantic model. Annotation-based models use relatively simple structure as annotations. Rich semantic model represents semantics in real-world manner with more complex structure.

The widely known MPEG-7 standard[3] mainly considers representation of low-level feature and syntactic information and strictly speaking it is not a data model for database since it does not provide query mechanism.

2.1 Annotation-based Models

The basic idea of annotation-based models is to put content information on top of video stream. Annotations may be predefined keywords, free text, or structured data. Each annotation is associated with a logical video segment. The relatively simple structure brings annotation-based models great flexibility, but also limits their expressive power and query supporting capability. Relations among annotations are not specified; abstractions of videos that do not appear in the video can not be modeled due to the tight-couple between annotations and segments' position; queries are barely supported as a declarative language, except only using keywords or attributes.

OVID[6], VideoStar[7] and CCM[8-9], which appears at early 90's, all utilize the fundamental annotation style, that is, add an annotation layer on top of video data and use attributes to describe semantics. OVID is a weak type model. No schema is needed. Each video object has its own attribute set and new attributes can be attached to it whenever necessary. Description data can be shared by "*interval-inclusion based inheritance*". VideoStar is a strong type model. Class *StoredVideoSegment*, *VideoStream* and *VideoDocument* are used to model physical video segment, logical video segment and the mappings between them respectively. The structure of a *VideoDocument* is represented by a hierarchy of *StructuralComponents*, each identifying a *FrameSequence*. Thematic indexing is supported by class *Annotations* which identifies a *FrameSequences* and gives a textual description of its contents. CCM is a compromise between strong type and weak type data model. It retains the property-inheritance mechanism of strong type models and at the same time provides flexible facilities for dynamic schema update as found in weak type models. Objects in CCM are dynamically aggregated into *clusters*; objects in a cluster can play various roles. Every cluster has its own attribute-set and method-set. Clusters can have subclusters. A subcluster may optionally inherit attributes/methods of its supercluster. Being dynamic constructs, (sub)clusters and roles can always be modified in terms of its attribute-set, method-set, and role-player association at any time.

At later 90's, other techniques are integrated into annotation-based models. Information retrieval (IR) was introduced by VideoText[10] which uses free text as annotations and use IR techniques to retrieve contents encoded in these free text. In its later work WVTDB[11], three layers of data abstractions (*logical video stream*, *logical video segment*, and *user view*) are built upon the *physical video clips* to achieve *physical video data independence*, *logical video data independence* and *user view independence*. Smart VideoText[29] is also an extension of VideoText in which conceptual graph is used to represent knowledge in free text annotations. The Strata-based approach proposed in [12] integrats video analysis techniques into data model. It segments video's contextual information into multiple strata. Each stratum describes the temporal occurrences of a simple concept such as the appearance of an anchor person in a news video. By performing video analysis, judiciously chosen strata can be extracted automatically. However, due to the state of arts of video analysis techniques, information that can be extracted this way is very limited.

2.2 Rich Semantic Models

With the improvement of users' requirement for video retrieval, annotation-based model can not satisfy users any longer, which caused the appearance of rich semantic models. Here by rich semantic models, we mean

those that have great power on expressing real word entities, such as concepts, objects, events, and relationships. These entities may be a concrete thing appearing in the video or an abstract concept, even those that do not appear in the video but function as background knowledge. Rich semantic models are richer than annotation-based models in two aspects. First, they can represent richer information, such as relationships between semantics. Second, they provide richer query ability such as query by attributes, relationships, temporal ordering, and browsing. This is why they are called rich semantic models. However, all things have two sides. Great expressive power also brings complexity. Thus, this kind of models usually is not as flexible as those annotation-based ones. When schema evolution occurs, great effort must be taken to deal with existing data.

Rich semantic models are usually layered, with the lowest layer corresponding to the raw video stream and the highest layer representing semantic information. Sometimes mid-layers are placed between them, such as logical video segment layer, feature layer, media object layer, etc. The representations of semantic information in these models are based on quite different strategies. One is to extend or utilize existing models to represent video semantics. The other one is to design from scratch.

VIMSYS[13,14] is one of the earliest data models for the management of images and has influenced many succeeding data models. So although it is a model for images, we include it here to be compared with other models for videos. There are four levels in VIMSYS: image presentation level, image object level, domain object level and domain event level. A distinctive concept in VIMSYS is domain event which connects different domain objects by spatio-temporal relationship.

Videx[17], Temporal OO Data Model[20], Extended $ExIFO_2$[24], Ahmet Ekin's model[27] and THVDM[28] take the first strategy.

Videx[17] uses UML notions to represent the structure and semantics of video data in an object-oriented manner. It integrates the low-level and high-level feature and abstracts the video data into two levels: logical video unit level and physical video unit level. A physical video unit may be a shot, a scene or a video sequence and can be related to multiple logical video units. A logical video unit contains domain-specific information so it can be easily extended to different domains.

Carlo Combi's Temporal Object-Oriented Data Model[20] adopts and extends an existing temporal data model named GCH-OODM to consider multimedia data. All classes modeling objects with temporal dimension are subclasses of class *Temporal_Object*. *Observation* is one of such subclass which is used to relate semantic information to video or part of video. This model explores the relationships among intervals that an observation being true, and divided observations into categories according to these relationships. Each of the categories corresponds to a subclass of *Observation*.

A model that can deal with uncertain and imprecise properties is proposed in [24] which is an extension of the data model $ExIFO_2$. The purpose of $ExIFO_2$ is to handle complex objects with their uncertain and imprecise properties. To meet requirements of multimedia application, new constructors such as *sequence* are added. Three levels of uncertainty are provided: attribute level, class/object level, and class/subclass level.

Ahmet Ekin's model[27] organizes video semantics based on events. It unifies the shot-based and object-based structural video models with the entity-relationship (ER) or object-oriented models. This model is an extension of ER models, with object-oriented concepts added. Entities in the model include Video Event and Video Object. The actor entity functions as relationships between Event and Object. All attributes of an object specific to an event are stored in the actor entity. Only event-independent attributes are stored in object entity. The schema and instance are all represented as a graph, so queries are evaluated by graph matching.

THVDM[28] is an integrated model to handle low-level feature information, syntactic information and semantic information. It has three layers corresponding to the three kinds of information respectively. Semantic information is described using a model named ERR which means ER model plus Rules. Rules are used to define events by means of objects and their spatio-temporal properties or relationships.

AVIS[15,16], VideoGraph[18,19], $CoPaV^2$[21], Semantic Associative Video Model[22,23], and BilVideo[25,26] take the second strategy.

AVIS[15] is the earliest rich video semantic model. It divides video into fixed-time duration frame sequences and uses a special kind of segmented tree called frame segmented tree to represent the structure of a video. Each node in the tree represents a frame sequence and the objects and events occuring in it. Activities are types of events. A set of arrays are used to store objects, events, activities and their associations. In their later work [16], the authors extended the model by defining feature-subfeature relationships. When a query cannot be answered, it can be relaxed by substituting a feature by its subfeature.

VideoGraph[18] and SemVideo[19] are proposed by the same authors with similar main ideas, both taking temporal relationships among semantic descriptions as components of the model and using them to infer implicit temporal information. The set of objects are divided into two classes, key objects and non-key objects. A key object has related temporal information telling what parts of the video is associated with it. Non-key objects have no

related temporal information. Temporal relationships between objects are represented by r-link. Thus, non-key objects' temporal information can be obtained by exploring the temporal relationship between it and related key objects. For each object, its component objects are connected to it by c-link.

In [21], a data model named CoPaV^2 and a rule based query language is developed for video indexing and retrieval. Two layers exist in this model: Feature & Content Layer and Semantic Layer. Entities in the Semantic Layer fall into three categories: objects and object identity, attributes, and relations. Two types of objects are considered. Temporal cohesion objects are abstract objects resulting from splitting a given video sequence into a set of smaller sequences; Semantic objects are entities of interest in a given video sequence. No object type hierarchy is provided, so each object can be viewed as a unique type with its own attributes. Relations among objects are stated explicitly as a first-class language constructs. Rules can be used to define new relations. Although CoPaV^2 is schemaless, conceptual model can be added by using rules to classify video data.

Yong put forward a Semantic Associative Video Model in [22,23], in which video's content is represented in three layers: scene layer, object layer and concept layer. Three data structures Scene Network, Semantic Object Net, and Hierarchical Concept Tree are used to store information in these layers respectively. Scene network is a network formed by scenes and temporal relationships between them. Semantic object net contains a set of shots, a set of semantic objects, a mapping function indicating the appearance of semantic objects in each shot, and a mapping function denoting relationships among semantic objects. The hierarchical concept tree (HCT) is a tree formed by abstract concepts and the ISA relationships among them. Semantic objects in semantic object net are also related to concept class by ISA relationship.

BilVideo[25,26] is a video database management system providing integrated support for spatio-temporal and semantic queries for video. The semantic data model in BilVideo has two layers: feature and content layer deals with low level details, and semantic layer deals with semantics perceived by human. Events, subevents, and objects are considered and form a hierarchy structure. Video consists of events and events consist of subevents. Objects are modeled in every level in the hierarchy. The model is mapped to relational tables for implementation.

3. Criteria and Evaluation of Rich Semantic Models

In Sections 2, we described eighteen video semantic models. To compare and better understand these models, in this section we will define comprehensive criteria for rich semantic models and evaluate the eleven rich semantic models using these criteria. Results are shown in Table 1 and 2 in chronological order.

These criteria are organized in rules and classified into three categories: the Expressive Power, the Acquisition of Semantic Information, and the Query Supporting Capability. These categories just fit in the three stages in developing video-related applications, that is, the stage of schema definition, data acquisition, and querying on established video database.

In following sections, these criteria will be presented along with the evaluation results of the eleven models. For evaluation, it is important to have a precise and complete definition of the models. However, descriptions of existing models range from very formal and detailed to very vague. We will try our best to overcome the difficulty to make the evaluation objective.

3.1 Expressive Power

Expressive power is the basic component of data model determing what can be represented by the model. It is the basis of query supporting. In this section, eleven rules E1, E2, E3 etc. are presented for evaluating the expressive power of a video semantic model. The results of the evaluation are given in Table 1 and Table 2 where columns from E1 to E11 indicate the evaluation result of every rule for the eleven models.

E1 — Support for Object and User Defined Object Attributes In the first part of the result of this criterion, the structure used to store objects are stated. From the table we can see that nearly all models use a specific structure Object to store object information, while two models VideoGraph and Yong's use slightly different structure. Besides, additional aspects objects related to the video should be described by user defined attributes. This fact is indicated by the keyword UDA. All models except AVIS support it.

E2 — Support for Event and User Defined Event Attributes In the first part of the result of this criterion the structure used to store events are stated. From the result of this rule and the previous rule, it's apparent whether a model treats object and event equally. The difference between object and event is that objects usually have a longer life span and attributes may take different values in different stages while events usually have a shorter life span and attribute values do not change. Thus there are two strategies to treat them: equally or differently. With equal treatment, uniform structure can be used to model both object and event, which will reduce the complexity of the model. On the other hand, although different treatment requires relatively more complex structure, it can provide more functionality regarding either objects or events.

From the table we can see that in general, later models treat them differently while earlier models treat them equally, among which Videx and Temporal OO are extension of object-oriented model and inherit the idea that "object is used to model everything". Similar to objects, events also need user defined attributes to describe different aspects of them. The same keyword UDA is used as the result. The result show that nearly all models support user defined event attributes. In AVIS, event attributes can only be defined as roles, so the UDA is put in parentheses. In the description of VIMSYS, nothing is stated about event attributes. We infer that user defined attributes are not supported since events in VIMSYS are computed from image sequences or videos resulting from motion, spatial interactions, appearance, disappearance etc.

Name	E1	E2	E3	E4	E5	E6	E7	E8	E9
VIMSYS [13,14]	Object, UDA	Event	No	Is_a(O)	No	(Yes)	No	No	No
AVIS [15,16]	Object	Event, (UDA)	Yes	No	Yes	No	No	No	No
Videx[17]	Object, UDA	Object, UDA	Yes	Is_a(O,E), temporal	No	Implicit	No	No	No
VideoGraph [18,19]	Internal Graph Node, UDA	Internal Graph Node, UDA	No	temporal	No	(Implicit)	SYS	No	Time
Temporal OO[20]	Object, UDA	Object, UDA	Yes	Is_a(O,E) , temporal	No	Implicit	No	Yes	Time
CoPaV2[21]	Object, UDA	Object, UDA	No	No	No	(Explicit)	SYS	No	No
Yong's [22,23]	Concept/Semantic Object, UDA	Concept/Semantic Object, UDA	(Yes)	Is_a(O,E)	No	(Explicit)	No	No	No
Extended ExIFO2[24]	Object, UDA	Event, UDA	Yes	Is_a(O,E)	Yes	(Explicit)	No	No	Attribute, Class/object, Class/subclass
BilVideo [25,26]	Object, UDA	Event, UDA	Yes	Event-subEvent	Yes	No	No	No	No
Ahmet Ekin's [27]	Object, UDA	Event	No	Is_a(O),Event-subEvent, causal, temporal	Yes	(Explicit)	No	No	No
THVDM [28]	Object, UDA	Event, UDA	Yes	Is_a(O,E), causal, temporal	Yes	Explicit	No	No	No

Table 1. Evaluation of Criteria E1-E9

E3 — distinguishing activity (event type) and event This rule examines whether activity and event are distinguished in a model. The relation between event and activity is like the relation between instance and class in Object-Oriented model. The advantage of distinguishing activity and event is that events can be managed by category and query about event's category can be answered directly instead of scanning all events. All models except VIMSYS, VideoGrpah, $CoPaV^2$, and Ahmet Ekin's model support this differentiation. In Semantic Associative Model, activities are nodes in hierarchical concept tree while events are semantic objects in semantic object net. However, how information of activity is managed and how queries about activities are processed is not stated. So the evaluation result for it is put in parentheses. In Videx, Temporal OO, Extended $ExIFO_2$, and THVDM, this is inherited from the base Object-Oriented model by means of class and instance.

E4 — Support for Special Relationships between Objects or Events This criterion evaluates to what extent a model supports special relationships between objects or events. These relationships are Is_a relationship between objects and events, event-subevent relationship, causal relationship, and temporal relationship. They are not related to specific domain and the awareness of them can help understanding of video content and serve users more

efficiently. The result lists relationships **explicitly** supported by each model.

With Is_a relationship, a hierarchical structure can be built for types of objects or events, which can help in many aspects such as: attributes can be inherited and reused; similarity between different types can be measured; queries can be reformulated along the hierarchical structure, etc. Most models support Is_a relationship between object types, while fewer support that between event types. For Videx, Temporal Object-Oriented Data Model, extended $ExIFO_2$, Ahmet Ekin's model and THVDM, this feature is supported as class-subclass relationship as in OO model. Semantic Associative Video Model uses the hierarchical concept tree to describe Is_a relationship between concept classes.

Event-subevent relationship is used to detail an event by describing its components (sub-events). This relationship is different from the Is_a relationship in that a sub-event is a part of an event, while a sub-class is a kind of an event. Two of these models, BilVideo, and Ahmet Ekin's model, have considered this relationship.

Temporal relationship between events is also an essential aspect for describing video content. Totally five models support temporal relationship as shown in table 1. For BilVideo and $CoPaV^2$, although temporal relationships are not represented explicitly, they can be used in query definition by using temporal comparison functions.

Causal relationships between events reveal an important aspect of a video. It helps to understand the semantics of the video and allows users to issue queries regarding the causal relationship. Among these models, only two support causal relationship between events: Ahmet Ekin's model and THVDM.

E5 — Objects' Roles This rule evaluates whether objects' roles when participating events are supported. These roles should be modeled **explicitly** instead of as events' attributes since they identify the interactions between objects and events. In general, earlier models do not have this feature while later models have, except AVIS.

E6 — User Defined Relationships between Objects or Events This criterion considers the modeling of user defined relationships between objects or events in a specific domain. The keyword "Explicit" and "Implicit" indicate whether relationships are modeled explicitly or implicitly as attributes. Most models support user defined relationships between objects and events and the modeling method has experienced the evolution from implicit to explicit. VideoGraph uses *c-link* to define relationships between objects (events) whose modeling power is limited by the no-circle-constraint; Semantic Associative Video Model and $CoPaV^2$ support only relationships between object instances, not object classes; Extended $ExIFO_2$, and Ahmet Ekin's model support only user defined relationships between objects, not events. To indicate these limitations, results for these models are parenthesized. For VIMSYS, relationships between objects can be defined, however, how they are represented is not stated, so a Yes in parenthesis is provided as the result.

E7 — Constraint Representation This rule evaluates the modeling power for constraints, including system constraints and user defined constraints. System constraints are used to constrain model structures while user defined constraints are used to describe domain-specific properties to ensure the validation of information. This feature is poorly supported in all existing models. Only two of them provide support for system constraints. VideoGraph poses two constraints on the structure of the graph; $CoPaV^2$ uses constraints to associate a time interval to a temporal cohesion object. As far as user-defined constraints are considered, no mechanism was mentioned in all these models.

E8 — Object History This criterion evaluates whether object history is supported. In E2 we have stated that an object usually has a long life span and during its life span values of its attributes may vary. An important thing is that it must be known that all these different values are describing the same entity. That is, we need a history to record evolution of an object. However, in these models only Temporal Object-Oriented Data Model has this ability. It uses the temporal functionalities provided by the temporal data model GCH-OODM to record history of objects of type Observation.

E9 — Support for Uncertainty This rule examines the ability of modeling uncertain information. The reason we need such a feature is that for video data, it is impossible or impractical to acquire complete semantic information. Besides, incomplete information may cause uncertainty in query results. So, in order to provide as much information as possible to users, uncertainty needs to be considered inherently by the model. The result of this criterion is the kind of uncertainty supported in each model. Totally three models considered uncertainty to different extents. In VideoGraph and Temporal OO, only uncertainty in time representation is supported. Extended $ExIFO_2$ supports three levels of uncertainty: attribute-level, class/object level, and class/subclass level.

E10 — Containing Low-Level or Syntactic Information This criterion considers the ability for modeling low-level and syntactic information. Although in this paper we concentrate on the semantic aspects of modeling, whether low-level or syntactic information is integrated is still an import aspect. One reason is that although infrequent, queries about low-level or syntactic information do exist. Another reason is that in the future, when semantics can be acquired automatically by video analysis, these algorithms can be integrated easily. The

keyword "Low" and "Syn" indicate low-level information and syntactic information respectively. From the table we can see that most models support one or two of them.

E11 — Modeling of Logical Video Segments This criterion evaluates whether physical data independency is provided. This is necessary since one physical video file does not always represent a logical meaningful segment. However, from Table 2 we can see that only Videx, Carlo Combi's Temporal OO Data Model, and Extended ExIFO$_2$ provide this ability.

Name	E10	E11	A1	A2	A3	A4	Q1	Q2	Q3	Q4	Q5	Q6
VIMSYS [13,14]	Low,Syn	No	No	No	No	No	S, IS	N.A.	Graphical Interface	Yes	No	No
AVIS [15,16]	No	No	No	No	No	No	No	N.A.	Procedural	No	Relaxation	No
Videx[17]	Low,Syn	Yes	No	No	No	No	S, T	N.A.	N.A.	No	No	No
VideoGraph[18,19]	No	No	No	No	Time	No	T	AND,OR,NOT IMPLY, EQ,UQ	Declarative	No	No	No
Temporal OO[20]	No	Yes	No	No	No	No	T	N.A.	N.A.	No	No	No
CoPaV2 [21]	Low	No	No	No	Defined Relation	No	T	AND, IMPLY	Declarative	No	No	Defined Relation
Yong's [22,23]	Syn	No	No	No	No	No	OS, SS,TL, B	N.A.	N.A.	No	No	No
Extended ExIFO2 [24]	Low	Yes	No	No	No	No	S,TL, U	N.A.	N.A.	No	No	No
BilVideo [25,26]	Syn	No	No	No	Spatial relations	No	T, S	AND,OR,NOT	Declarative	No	No	Spatial Relationship
Ahmet Ekin's [27]	Syn	No	No	No	No	No	T, B	N.A.	Graphical Interface	No	Relaxation	No
THVDM [28]	Low,Syn	No	No	No	No	No	S, T,A	AND,OR,NOT	Declarative	No	No	No

Table 2. Evaluation of Criteria E10,E11,A1-A4, Q1-Q6

3.2 Acquisition of Semantic Information

A major issue concerned with video application is how to acquire semantic information for the schema of selected models. We call it problem of information acquisition. In traditional relational database, information acquisition is not a big issue. But for video semantic model, information of objects, events, spatial or temporal relations etc. are related to understanding of videos. A well designed video semantic model may provide support to facilitate the acquisition of semantic information. Criteria in this section are presented to examine this capability.

A1 — Encoding Domain Knowledge This criterion evaluates whether domain knowledge can be integrated when defining a schema for a specific domain. Since for a specific application, videos in the database always have a uniform subject, if domain knowledge can be encoded in the process of schema design, they can help a lot for semantic information acquisition, analysis and query evaluation. These domain knowledge may take various forms such as constraints for attribute values, video structures, temporal relationships, or aspects considered when design model components. Currently no model provides this capability.

To acquire semantic information, two steps should be taken: first an initial set of semantics is got through video analysis or annotation, then inference may be performed to augment it and get an extended semantic information set.

A2 — Helping the Acquirement of Initial Semantic Information This criterion measures the ability of facilitating the first step of semantic information acquisition. Initial semantic information may be acquired by video analysis or annotation. For annotation, users usually first form a logical video segment and then input semantics or select existing semantics and relate them to the logical video segment. If domain knowledge about the structure or subject of the video is known, hints may be

given to help users to find a logical video segment or an existing semantics. This feature is also poorly supported.

A3 — Inferring New Information from Existing Information This criterion measures the ability of facilitating the second step of semantic information acquisition. Due to the complexity of video content and different views of users, enormous things are implied in a video. Annotating all these things is usually impossible or impractical. As the result, if new information can be inferred from existing information, users can get more information with the same workload which can also be viewed as the reduction of annotation workload. The result of this criterion is the kind of information that can be inferred. Three models support inference of information to different extent. In VideoGraph, non-key objects' temporal information can be obtained by considering temporal information of related key objects and temporal relationships between them. In CoPaV^2, new relations can be defined using existing relations, so the newly defined relation can be inferred given existing relations. In BilVideo, new spatial relations can be inferred based on known spatial relationships stored in the fact base.

A4 — Detecting Logical Errors in Semantic Information This criterion deals with errors occurring in the two steps of acquiring semantic information. Human annotator, video analysis algorithm, and inference algorithm are not absolutely correct and may cause logical errors in the semantic information acquired. For example, the first section of a game is labeled as the second section and the second one labeled as the first one. If rules can be defined to ensure that the first section should occur before the second one, this kind of mistakes can be detected and avoided. However, since no models provide the mechanism to define this kind of constraints, no capability of detecting errors is provided.

3.3 Query Supporting Capability

The ultimate purpose of video database is to provide query service for users to find interested data. Thus, query supporting capability is an important aspect subject to evaluation.

Q1 —Types of Queries Supported This criterion evaluates to what extent some special kinds of queries are supported in each model. In Section 3.1, we put forward several criteria to evaluate the expressive power. Usually information that can be represented can also be used to issue queries. So the ability evaluated by this criterion is about some special kinds of queries, including spatial(S), temporal(T), aggregate(A), uncertainty (U), query by object similarity(OS) and event similarity(ES), query by shot similarity(SS), and browsing(B). Spatial query is query about salient object, their spatial location(SL), and spatial relationships(SR). Temporal query refers to query about temporal location(TL) and temporal relationship(TR). Aggregation query is like the GroupBy clause in SQL, returning information about a group of objects or events. Uncertainty query is query involving uncertainty in its definition or result. Browsing is an important and useful functionality, by which users can have an overview of the whole database or a specific video. The result of this criterion is the kinds of queries supported. From table 2 we can see that nearly all models support temporal query; spatial query is support by most models; query about object similarity and shot similarity is supported by semantic association model; uncertainty query and aggregation query are supported by only one model, Extended ExIFO$_2$ and THVDM respectively; the browsing query mode is only provided in two models: semantic association model and Ahmet Ekin's model.

Q2 — Constructors for Query Condition This criterion considers what constructors can be used in each model to construct query conditions which decides how complex the query may be. Possible connectors maybe Boolean operators (including AND, OR, NOT), and logical constructors (including IMPLY, existential quantifier (EQ), and universal quantifier (UQ)). The result of this criterion is the constructors supported.

Q3 — Query Language This criterion considers what kind of query language is provided. The form of the language decides how friendly it is, which may be declarative, procedural or graphical interface. When the effort taken by users to learn to use the query language is considered, the order of the three forms from easy to difficult is: graphical interface, declarative, and procedural. The result of this criterion is the form of the language provided in each model. Except AVIS, most models provide at least declarative language.

Q4 — Support for Query Evolution This rule evaluates the ability of supporting query evolution. This is necessary since in video applications sometimes users do not have a clear idea about what he wants to get. He may first randomly issue a query and browse the results, and refine the query when he sees something interesting. This requires that query can be issued over the result of previous queries. Among existing models, only VIMSYS has this ability.

Q5 — Support for Query Reformulation This rule considers whether query reformulation is supported. In video databases, due to the incompleteness and uncertainty, when empty result occurs, the reason may be the imprecise in query definition or query evaluation process. In this situation, a model is expected to adaptively provide something that may be useful for users, that is, relax the query and return a superset. In another situation, when too many results are returned, to make it more informative, the query condition should be strengthened to reduce the size of the result set. Two models, AVIS and

Ahmet Ekin's model, support query relaxation. In AVIS, features can be substituted by sub features to relax a query condition; in Ahmet Ekin's model, partial matching of query condition can be performed. However, support for query strengthens is not mentioned in all existing models.

Q6 — Support for Inference This criterion evaluates whether inference can be made during the process of query evaluation. In CoPaV^2, rules are used to define new relations using existing relations. When evaluating queries, with the knowledge of existing relations, these newly defined relations can be inferred and evaluated. In BilVideo, spatial relationships can be inferred from the fact base. The result of this criterion is the information that can be inferred when evaluating a query.

4. Learning from the Evaluation Results

In section 3 we have proposed totally 21 criteria for video semantic model and evaluated existing typical rich video semantic models according to these criteria. In this section we will analyze the evaluation results to see what we can learn.

As far as consideration of individual criterion is concerned, some columns in Table 1 and 2 show apparent evolving path along the development, such as modeling of objects, events, object roles, user-defined relationships, and query language (corresponding to column "E1", "E2", "E5", "E6" in Table 1 and column "Q3" in Table 2 respectively). These trends not only show evolving process of existing models, but also give directions to future models.

From the column "E1" and "E2" in Table 1, we can see that most models support modeling of objects, events and their attributes. However, these models are different on whether objects and events are treated differently. In the description of criterion E2 we have stated the difference of objects and events and we can know easily whether a model treats them differently by comparing the two columns. As we can see, earlier models usually treat them equally, such as Videx, VideoGraph, Temporal OO Data Model, CoPaV^2, and Semantic Associative Model, while later models treat them differently, such as Extended ExIFO$_2$, BilVideo, Ahem Ekin's model and THVDM.

The column "E5" in Table 1 shows whether each model represent objects' roles in events explicitly. The first video semantic model AVIS represents roles explicitly. However, several models after it, including Videx, VideoGraph, Temporal OO Data Model, CoPaV^2, and Semantic Associative Model treat roles as events' attributes. In this manner, it is unknown whether an attribute of an event describes the event itself or the interactions between it and other objects. Thus, all later models, including Extended ExIFO$_2$, BilVideo, Ahem Ekin's model and THVDM, take roles as an explicit model component.

The column "E6" in Table 1 shows whether user-defined relationships are supported in each model and whether they are represented explicitly or implicitly. The development process is from "implicit" to "explicit". In Videx, VideoGraph and Temporal OO Data Model, relationships are modeled implicitly as attributes or components. As many researchers argued, this method sometimes can not model the real world in the most natural way. So in later models such as CoPaV^2, Semantic Associative Model, Extended ExIFO$_2$, BilVideo, Ahem Ekin's model and THVDM, relationships are modeled explicitly.

The column "Q3" in Table 2 shows the query language provided by each model. Except the first video semantic model (AVIS) who provides only procedural query interface, most models provide declarative query language or graphic interface to serve users more friendly.

As far as overall consideration of all criteria is concerned, aspects considered by each model also change along the development. AVIS is the earliest rich semantic model for video. It has many notable advantages superior to earlier annotation-based models: clear separation of objects and events (column "E1" and "E2" in Table 1); managing events by category (activity) (column "E3" in Table 1); and using ROLE and PLAYERS to relate objects to events (column "E5" in Table 1). This kind of generalization is very close to human's perception of the real world. However, AVIS also has some drawbacks, such as using fixed-duration frame sequences as a unit to relate semantic information; no relationships between event and object except Participation is allowed; attributes are only available as events' roles; and only a procedural querying facility is provided.

In later models, some drawbacks of AVIS are mended. The most remarkable improvements are in expressive power and query ability. For expressive power, almost all later models support user defined attributes of objects and events (column "E1" and "E2" in Table 1); most models distinguish event and event class (column "E3" in Table 1); user defined relationships between objects and events are supported implicitly or explicitly (column "E5" in Table 1); and shots are used as video segments instead of fixed-duration frame sequences. For query ability, they allowing users to define more complex query conditions (column "Q2" in Table 2), provid declarative query languages or graphical interfaces (column "Q3" in Table 2), and provid more querying modes such as spatial query, aggregation query etc.(column "Q1" in Table 2). Some new features are added, including support for special relationships such as Is_a relationship, temporal relationship, causal relationship, and event-subevent relationship (column

"E4" in Table 1), object history (column"E8" in Table 1) and uncertainty (column "E9" in Table 1).

Although great improvements have been made, there are still important advanced features that are supported poorly or even not supported at all. For expressive power, causal relationship, event-subevent relationship (column "E4" in Table 1), object history (column "E8" in Table 1) and uncertainty (column "E9" in Table 1) are supported only in a few models with a very limited extent. Event-subevent relationship is supported only by BilVideo and Ahmet Ekin's model while in BilVideo it is a one-level relationship which means sub events may not have sub events. Object history is supported only by Temporal Object-Oriented model. Uncertainty is supported by VideoGraph, Temporal OO, and Extended ExIFO$_2$, while VideoGraph and Temporal OO only support uncertainty in time representation. As far as user-defined domain-specific constraints is concerned, unfortunately no current model provides mechanisms to support it (column "E7" in Table 1). Only a small fraction of models support logical video segment modeling (column "E11" in Table 2). For query ability, support for inference (column "Q6" in Table 2) in query evaluation, query evolution (column "Q4" in Table 2), and query reformulation (column "Q5" in Table 2) are very limited. Query modes provided are not very powerful (column "Q1" in Table 2). The significant browsing mode is only supported by two models (Semantic Associative Model and Ahmet Ekin's model) with a limited browsing granularity. In Semantic Associative Model, users can only browse videos by scenes, while in Ahmet Ekin's model users can browse videos by entities. Query by uncertainty, object similarity, event similarity or shot similarity and aggregation query service are provided by few models. When considering utilizing the model to help the process of semantic information acquisition (column "A1", "A2", "A3", "A4" in Table 2), nearly nothing is provided by existing models. However, these aspects are all inevitable for a model to serve a specific domain.

As far as the two strategies of designing a model (mentioned in section 2) is concerned, no apparent trend can be found. The uses of the two strategies are interleaved. Thus when designing a new model, designers should choose the strategy according to specific situations.

Another overall consideration for a video semantic model is whether the model should be general or specialized. The fact is that nearly all existing models are general. However, the generality has limited the utilization of properties of videos in a specific domain, thus requirements concerned with these properties cannot be supported by these models. In the other extreme, specialized models may be well designed powerfully for a certain domain but cannot be used in other domains. As a result, we have to achieve a good balance between the generality and powerfulness of video semantic models.

5. Conclusion

A survey of existing video semantic models is presented in this paper. These models are classified into two categories: annotation-based models and rich semantic models. To evaluate these models and to help users to choose or design appropriate models, we present twenty one rules. These rules cover the whole process of model-based video application development. According to these criteria, eleven typical rich semantic models are evaluated. The evaluation results show apparent evolving path in some aspects such as modeling of objects, events, object roles, user-defined relationships, and query language. The results also show that most efforts are put on basic expressive power and basic query ability. Some advanced features are poorly supported. Representation of user-defined domain-specific constraints and facilitation of acquisition of semantic information are seldom considered. Furthermore, to avoid repeated work of applying a model to different applications, semantic model designed for a class of domains with common features is needed.

References

1. Aas, K.; Eikvil, L. A Survey on: Content-based Access to Image and Video Databases. Report 915, Norwegian Computing Center, March 1997.
2. Pekovic, M.; Jonker, W. An Overview of Data Models and Query Languages for Content-based Video Retrieval; proc of International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet 2000
3. José M. Martínez. Overview of the MPEG-7 Standard. ISO/IEC JTC1/SC29/WG11 N6828, October 2004. http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm
4. Bashir, F. I.; Khokhar, A. Video Content Modeling Techniques: An Overview; 2002
5. Spatiotemporal Data Modeling and Management:A Survey; 1999 , http://www.itee.uq.edu.au/~zxf/_papers/STDBSurvey.pdf
6. Eitetsu Oomoto; Katsumi Tanaka. OVID: design and implementation of a video-object database system. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 1993, pp. 629-643
7. R. Hjelsvold and R. Midtstraum. (1994) Modeling and Querying Video Data. Proc. 1994 Intl. Conf. on Very Large Databases, pp. 686-694 Santiago. Chile.
8. Li, Q.; Lee, C.M. Dynamic object clustering for video database manipulation. In proceedings of IFIP 2.6 Working Conference Visual Database Systems (VDB-3), 1995
9. Qing Li A dynamic data model for a video database management system. ACM Computing Surveys 1995, 27, 602-606

10. H.T. Jiang, D. Montesi, and A. K. Elmagarmid (1997). VideoText database systems. In Proceedings of IEEE International Conference on Multimedia Computing and Systems, pp 344-351.
11. Haitao Jiang; Ahmed K. Elmagarmid. "WVTBD- A Semantic Content-Based Video Database System on the World Wide Web". IEEE Transactions on Knowledge and Data Engineering, 1998: 10(6), 947-966
12. Kankanhalli, M. S.; Chua, T.-S. Video modeling using strata-based annotations. IEEE Transactions on Multimedia 2000, 7, 68-74
13. Gupta, A.; Weymouth, T.; Jain, R. Semantic Queries with Pictures:The VIMSYS Model; VLDB 1991; pp 69-79
14. Data Model and Semantics in Visual Information Management Systems, http://www.cs.wisc.edu/~beechung/dlm_image_processing/image_processing/vimsys.html
15. S. Adali; K.S. Candan; K. Erol; V.S. Subrahmanian ; AVIS: An Advanced Video Information System, technical report, Presented at the First International Workshop on Multimedia Information Systems 1995
16. Adah, S.; S.Candan, K.; Chen, S.-s. The Advanced Video Information System: Data Structures and Query Processing. ACM Multimedia Systems 1996, 4, 172-186
17. Tusch, R.; Kosch, H.; Boszormenyi, L. VIDEX: An Integrated Generic Video Indexing Approach; ACM Multimedia, 2000; pp 448-451
18. Tran, D. A.; Hua, K. A.; Vu, K. VideoGraph:A Graphical Object-based Model for Representing and Querying Video Data; 2000; In Proc. of the 19th International Conference on Conceptual Modeling (ER2000), pp 383-396
19. D. A. Tran; K. A. Hua; K. Vu. Semantic reasoning based video database systems. In Proc. Of 11th International Conference on Databases and Expert Systems Applications, London, U.K, September 2000
20. Combi, C. Modeling temporal aspects of visual and textual objects in multimedia databases; International Workshop on Temporal Representation and Reasoning 2000; pp 59-86
21. Hacid, M.-S.; Decleir, C. A database approach for modeling and querying video data. IEEE Transactions on Knowledge and Data Engineering 2000, 12, 729-750
22. Yong, C.; De, X. Hierarchical semantic associative video model; In proc of IEEE International Conference on Neural Networks and Signal Processing, 2003; pp 1217-1220
23. Yong, C.; De, X. Content-based semantic associative video model; In proc of International Conference on Signal Processing, 2002; pp 727-730
24. Aygun, R. S.; Yazici, A. Modeling and Management of Fuzzy Information in Multimedia Database Applications; Technical Report, 2002, State University of New York, USA
25. Arslan, U. A Semantic Data Model and Query Language for Video Databases; Master thesis, 2002, Bilkent University
26. Donderler, M. E. Data Modeling And Querying For Video Databases; Phd thesis, 2002, Bilkent University
27. A. Ekin, Sports Video Processing for Description, Summarization, and Search (Chapter 2 Structural and Semantic Video Modeling), phd thesis, http://www.ece.rochester.edu/users/tekalp/students/ekin_thesis.pdf , 2003
28. Yu Wang, Chunxiao Xing, Lizhu Zhou, THVDM: A Data Model for Video Management in Digital Library, proceedings of the 6th International Conference of Asian Digital Libraries, 2003; pp. 178-192
29. Kokkoras, F.; Jiang, H.; Vlahavas, I.; Elmagarmid, A.; Houstis, E.; Aref, W. Smart VideoText: a video data model based on conceptual graphs. Multimedia Systems 2002, 8(4), 328-338

Yu Wang She received the B.S. and M.S. degrees in Computer Science from Xi'an Jiaotong university, China in 1999 and 2002 respectively. During 2002-2006, she stayed in Department of Computer Science and Technology of Tsinghua University, China as a PHD candidate.

ChunXiao Xing professor of the department of computer science and technology in Tsinghua University, China. His research direction is distributed multimedia information system, high performance computer network, multimedia database and digital library.

LiZhu Zhou professor of the department of computer science and technology in Tsinghua University, China. His research area includes information system, database system, digital library, ontology-based search and representation, etc. He is now commissary of national technology committee of China, national computer science education direction committee of China, national computer science institute of China, and national database committee of China. He is also members of steering committee of PAKDD. He has been the chairman of the program committee of ICIPS in 1997, PAKDD in 1999, and International conference on New Information Technology in 2001.