A Semi-Automatic Semantic Annotation Tool for Video Databases

Umut Arslan, Mehmet Emin Dönderler, Ediz Şaykol, Özgür Ulusoy and Uğur Güdükbay

Department of Computer Engineering, Bilkent University
06533 Bilkent, Ankara, Turkey
{aumut,mdonder,ediz,oulusoy,gudukbay}@cs.bilkent.edu.tr

Abstract. Advances in compression techniques, decreasing cost of storage, and high-speed transmission have facilitated the way video is created, stored and distributed. As a consequence, video is now being used in many application areas. The increase in the amount of video data deployed and used in today's applications not only caused video to draw more attention as a multimedia data type, but also led to the requirement of efficient management of video data. Management of video data paved the way for new research areas, such as indexing and retrieval of videos with respect to their spatio-temporal, visual and semantic contents. In this paper, semantic content of video is studied, where video metadata, activities, actions and objects of interest are considered within the context of video semantic content. A data model is proposed to model video semantic content, which is extracted from video data by a video annotation tool. The work in this paper constitutes a part of a video database system to provide support for semantic queries.

1 Introduction

Advances in compression techniques, decreasing cost of storage, and high-speed transmission have facilitated the way video is created, stored and distributed. These improvements created new application areas, where large amounts of video data are used, such as digital libraries, public information systems, video-on-demand systems, e-commerce, etc. More and more videos are created each day and this leads to an enormous growth in the number of videos to be dealt with. The fast increase in the amount of video data caused video to draw more attention as a multimedia data type and also revealed an important problem; new methods should be developed to manage it because existing data management techniques do not provide sufficient support for video.

Compressed video streams are examined to annotate motions of objects that appear in video. Automatic feature extraction techniques cannot directly extract semantic information from videos, but a number of

systems have been proposed that model high-level data like events in video. However, these systems are generally domain specific (e.g., news and sports) and cannot be used to model every type of video. Besides, objects appearing in videos can be annotated by the help of object extraction algorithms [11].

In this paper, we propose a semantic video model in which, video is modeled in a hierarchy of *events*, *subevents* and *objects* of interest. A video consists of events and an event consists of subevents. Moreover, objects are modeled in every level in the hierarchy. An event is an instance of an activity, which may involve many different objects over a time period. Subevents are used to detail an activity (event) into actions, and to model relations between objects of interest. The hierarchical model provides many semantic levels that facilitate understanding of video content. We have constructed a database model to have proper database management support for the semantic video model. We have also implemented an annotation tool to extract semantic information from videos, and to view and update semantic information that has already been extracted.

The work stated in this paper constitutes a part of *BilVideo* video database management system [6, 7] to provide support for semantic queries. BilVideo includes a rule-based spatio-temporal model for videos and a video query processor, which can answer spatial, temporal, similarity-based object trajectory, trajectory projection queries for videos. The organization of the paper is as follows: Section 2 discusses the related work on semantic querying systems. The semantic video model is presented in Section 3 and Section 4 presents the annotation tool. Finally, Section 5 concludes the paper.

2 Related Work

In the literature, there are numerous works about indexing, modeling, and retrieval of the semantic content of videos. As stated in [3], semantic conceptualization can be performed at several levels of information granularity. At the finest level of granularity, video data can be indexed based on low-level features such as color, texture, shape, and objects. At a coarser level of granularity, indexing of video data can be focused on activities, actions which are higher level abstractions. Automatic indexing of video data is desirable since manual indexing is hard and indexes that are created may differ with respect to the indexers. Low-level features, which can be extracted from video data without user intervention, have been used in automatic indexing of video data [2, 10]. However, low-level

features are not sufficient enough to index video data based on higher level abstractions.

A spatio-temporal model is proposed in [5] to model semantic information of video. Modeling events by using spatio-temporal attributes of objects is performed but this can only be used for specific domains like sports videos. A 'pass event in a soccer game' can be modeled by using spatio-temporal attributes but a 'party event' cannot be modeled in this way.

In [1], a semantic video model is proposed and the algorithms for handling different types of queries are implemented within a prototype, called Advanced Video Information System (AVIS). In this model, video is divided into fixed-time duration frame sequences. Activities, events and objects are related to the frame sequences and these relations are modeled by using a frame segment tree and arrays that store activities, events and video objects. Dividing video into fixed-size time intervals is not a good solution for temporal modeling of video. The query language proposed cannot answer temporal queries.

In [9], Common Video Object Tree (CVOT) model is proposed and video is modeled using spatial attributes of objects. In this model, all common objects among video clips are found and video clips are grouped according to these common objects. This data model is integrated into a temporal object model to provide concrete object database management support for video data. Temporal attributes for events and objects are supported by storing history of events and objects. Semantic attributes for objects and roles for activities are not addressed by the system.

In [8], a data model and a rule-based query language is developed for video content-based indexing and retrieval. The data model is designed around the object and constraint paradigms. The data model consists of *feature and content layer* and *semantic layer*. The semantic layer includes objects, attributes of objects, and relations between objects. The query language can be used to infer relationships about information represented in the model. Queries can refer to both of the layers.

3 Semantic Video Model

Modeling is necessary for efficient management and retrieval of videos. Semantic video modeling is the translation of video data into an internal representation, which captures the semantic content of video and creates indexes for efficient retrieval.

Video has two layers; *feature and content layer* that deals with the low level details of video, and *semantic layer* that deals with the meaning perceived by humans from a video. A semantic video model should capture events, subevents, objects of interest and bibliographic data about video. Actions are the acts performed by living objects. Data that is related to video itself, such as name, year of production, producer and etc., is specified as bibliographic data about video.

3.1 Hierarchical Structure

A video is modeled as a hierarchy of events, subevents and objects of interest. A hierarchical model provides many semantic levels that facilitate understanding of video content. Video consists of events and events consist of subevents. Moreover, objects are modeled in every level in the hierarchy. In the semantic video model, segmentation of video into sequences and scenes is performed by specifying events and subevents of video since events are associated with sequences and subevents are associated with scenes.

3.2 Data Model

Video consists of events. Events are the instances of activities taking place in video. In other words, activities are the abstractions of events. For example, wedding is an activity, but wedding of Richard Gere and Julia Roberts in a movie, is an event. Activities can be thought of as classes, and events can be thought of as the instances of these classes. For each activity type, a number of roles are defined. For example, murder is an activity. Murder activity has two roles defined for it: *murderer* and *victim*. The murder of Richard Gere by Julia Roberts is an event where Richard Gere has the role 'victim' and Julia Roberts has the role 'murderer'.

Subevents are used to detail events and to model the relations between objects of interest. To clarify the difference between events and subevents, assume that a party is depicted in a video. The party is modeled as an event that may contain a number of subevents: drinking, eating, dancing, talking. Several objects of interest can take place in this party event. These objects are assigned roles, which may be defined as 'host' and 'guest' for the party event. Actions represented by subevents, such as drinking or eating, are performed by living objects. These imply that objects are not only assigned roles defined for the event, but also they are associated with subevents, where they perform the actions represented by subevents.

Video name, duration, producer, director, video type, audience and subject of video are classified as bibliographic data. The attributes of interesting objects and values for the attributes are stored in object data whereas data related to events and subevents is stored in event data. Type of activity, begin and end times, objects that take part in an event, roles for objects, location and time are described as event specific data. Subevent specific data is given as follows: type of subactivity, begin and end times, and objects that appear in a subevent.

To sum up, events, subevents, objects, and bibliographic data form the abstraction of video semantic content. A database model is required to have proper database management support for the semantic video model. Regarding the specifications of the semantic video model, a database is created to store the semantic data of videos (see Figure 1). Detailed discussions on conceptual design of the database can be found in [4].

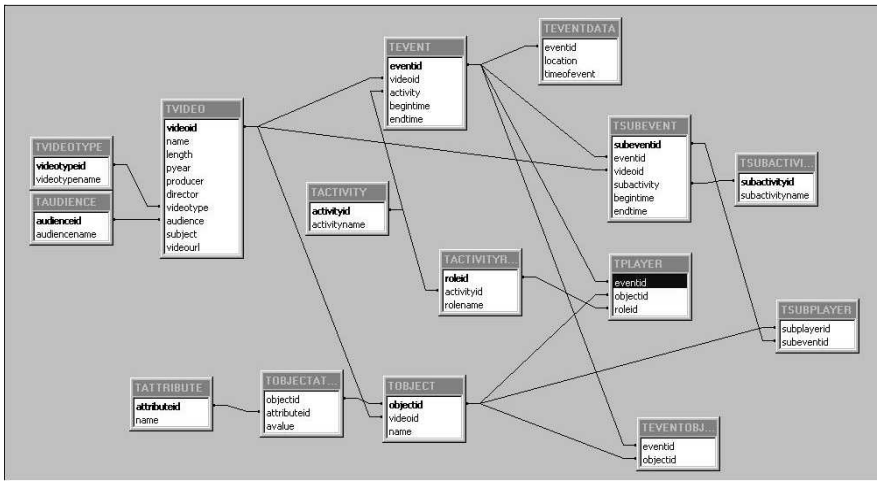

Fig. 1. Database Design of the Semantic Video Model.

3.3 Temporal Management

Temporal management of video segments can be categorized into three groups: *segmentation*, *stratification* and *temporal cohesion*. Segmentation splits video into independent and contiguous time segments, which allows one level of segmentation to be specified. Stratification allows overlapping of time segments, which provides many levels of segmentation to be performed. In temporal cohesion, a time segment is defined as a set of

non-overlapping time intervals and this provides many levels of segmentation and accurate representation of video segments. Temporal cohesion, which allows accurate temporal representation of time segments, is used in our semantic video model. Events and subevents are the time segments in our model. Video consists of events, which may overlap. Events consist of subevents, which may not be contiguous in time. Scenes may also overlap. These features provide flexibility in modeling activities and actions in video.

4 Video Annotator Tool

The video annotation tool is developed for annotating video clips according to the semantic video model. The tool is also used to view, update and delete the semantic data that has been extracted before. Semantic data extracted from a video may be categorized into five groups as follows:

1. *Metadata about a video:* Metadata contains the video specific data, such as video name, length of video, year of production, etc.
2. *Object data:* Object data is formed by items of interest in video.
3. *Event data:* Data related to activities that take place in video is considered as event data.
4. *Subevent data:* Data related to actions that take place in activities is considered as subevent data.
5. *Utility data:* Utility data consists of audiences, video types, activities, activity roles, sub-activities and object attributes.

4.1 Hierarchical Annotation Order

The order of annotation should follow the hierarchical semantic model of video from top to bottom. Hence, video is annotated first according to the hierarchy. Annotation of events with their corresponding subevents may be accomplished afterwards. During the annotation process, annotation of objects may be carried out whenever needed. However, the annotator must comply with the following restrictions:

- Utility data annotation is required for video metadata, event, subevent and object annotations.
- Event and object annotation cannot be done before video metadata annotation.
- Event annotation cannot be done before the annotation of objects of that event.

– A subevent cannot be annotated before the annotation of the event with which it is associated.

The annotation of utility data can be done at any time. However, utility data is required for the annotation of video metadata, events, subevents and objects. For example, video type and audience information are required during video metadata annotation. Since event and object annotations depend on video metadata annotation, event annotation cannot start immediately after video metadata annotation since for an event annotation to be complete, with event objects and roles of the objects specified, annotation of objects in the event must be done before the annotation of that event. Subevent annotation must be associated with an event annotation as well; if the event annotation is not done, then subevent annotation is not possible.

4.2 Hierarchical Video Tree

Video is modeled as a hierarchy in the semantic video model. The hierarchical video tree is used to show the current annotation status. The following rules define the hierarchical video tree:

– Root of the tree is a video entry.
– The leaves of a video entry are events and video objects.
– The leaves of an event are event objects and subevents.
– The leaves of a subevent are subevent objects.
– The leaves of a video object are the attributes and their values.
– The leaves of an event object are the roles of the event object in the activity.
– Subevent objects have no children.

4.3 Annotation Process

The annotation process is performed as follows: first of all, in the utility window shown in Figure 2, video types, audiences, object attributes subactivity types, activity types and roles for activity types are determined. Semantic annotation of a video starts with video metadata annotation, and video specific data is annotated. Video length is automatically retrieved from video player. 'Video Type' and 'Audience' fields can be selected from a list of choices. The 'new' button, when clicked, shows the 'Utilities Window', where utility data is annotated. For an event annotation, the objects that appear in that event should be annotated first.

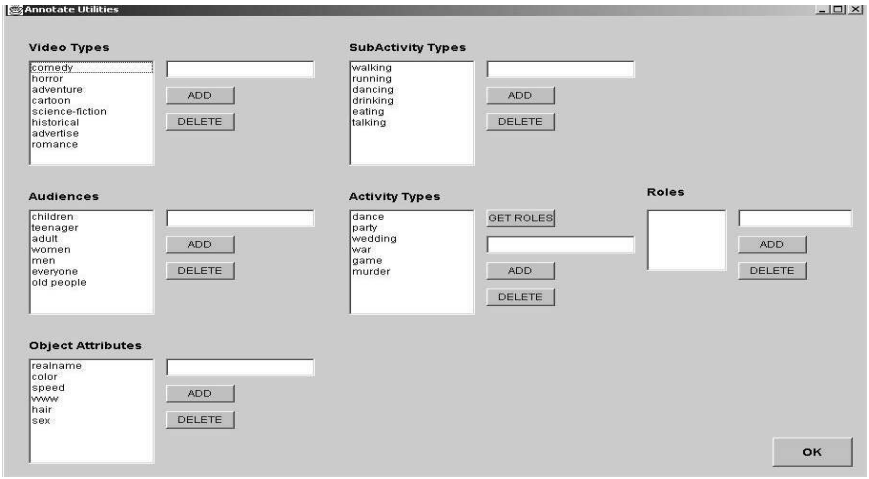

Fig. 2. Utility Window.

Objects can be added to and deleted from the video. Attributes for objects and values for the attributes can be defined or deleted. Only one value can be defined for each attribute.

The next step following the object annotation is the annotation of an event. For an event annotation, three windows are used: *event specification*, *object selection*, and *role definition*. In the event specification process, several attributes of an event is specified. To set the time interval, the video player is used directly, and by pausing the player, the time interval of an event is set. Event type is selected from a list of choices. Roles of the selected event can be retrieved and listed. New event types (activities) and roles can be defined in the utilities window. In object selection process, the video objects selected from the list form the objects to appear in the event. In role specification, video name and event type are given, and roles defined for the event type (activity) are also displayed together with the objects appearing in the event. The annotator has to match the objects with roles. One object may be associated with more than one role in the event.

Subevent annotations are performed last. Video name and event type are displayed to provide information to the annotator. Event type is used to show into which event the subevent is being inserted. Subevent type is selected from a list of choices, and the 'new' button is used to define new subevent types. Begin and end times are specified in the same way as it is done in the event annotation. The hierarchy after the annotations

is displayed in Figure 3. Subsequent annotations of other events should follow the same order of annotation (object → event → subevent). The order of upper group of buttons in the right side of the main window from top to bottom also reflects this order of annotation.

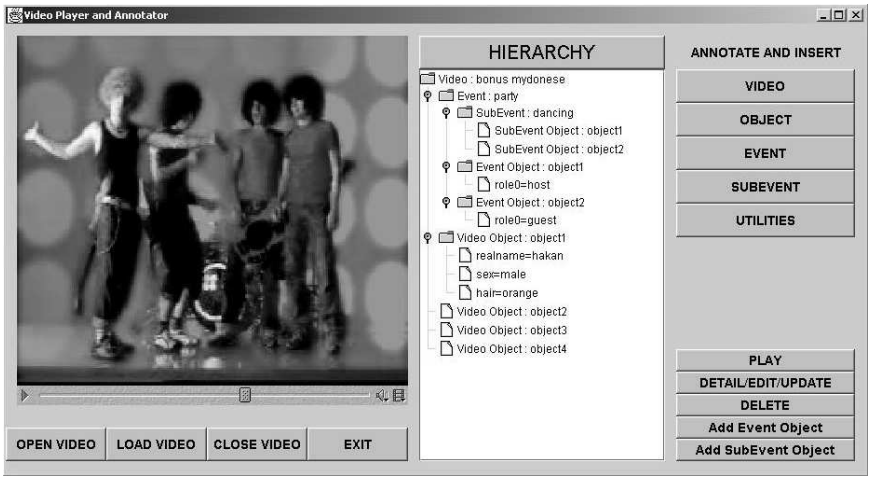

Fig. 3. The Hierarchy after Annotation Process.

5 Conclusion

Video has its own characteristics that differentiate it from other types of data. Management of video covers modeling, indexing and retrieval of video data. In this paper, we have looked at the management of semantic information of videos. We have defined the content of the semantic information to include bibliographic data about video and events, actions, and objects of interest taking place in video.

We have proposed a semantic video model which models video semantic information in a hierarchy. A video consists of events, and an event consists of subevents. Objects are modeled in every level in the hierarchy. A hierarchical model provides many semantic levels that facilitate understanding of video content. Temporal cohesion approach has been used to model time segments of video, which provides flexibility and accuracy in modeling events and subevents. A database model has been constructed to have proper database management support for the semantic video model.

We have implemented an annotation tool in Java to extract the semantic information from videos. Since manual annotation of video content

is a tedious process, extraction of information automatically is desirable. However, automatic information extraction techniques are not powerful enough to model video semantic content. Thus, human assistance is required in modeling video semantic information. The tool enables the annotator to see the current status of annotation in a hierarchical tree abstraction. The annotation tool simplifies the manual annotation process by providing simple and easy-to-understand user interfaces.

We are currently working on semantic query execution within BilVideo using the information within the semantic video model as a result of the annotation process. By the help of this model, semantic queries including event, subevent, and object conditions as well as video specific data queries will be executed.

References

1. S. Adali, K.S. Candan, S. Chen, K. Erol, and V.S. Subrahmanian. Advanced video information system: Data structures and query processing. *ACM-Springer Multimedia Systems Journal*, 4(4):172–186, 1996.
2. G. Ahanger and T.D.C. Little. Data semantics for improving retrieval performance of digital news video systems. *IEEE Transactions on Knowledge and Data Engineering*, 13(3):352–360, 2001.
3. W. Al-Khatib, Y.F. Day, A. Ghafoor, and P.B. Berra. Semantic modeling and knowledge representation in multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):64–80, 1999.
4. U. Arslan. A semantic data model and query language for video databases. M.S. Thesis, Department of Computer Engineering, Bilkent University, Ankara, Turkey, January 2002.
5. Y.F. Day, A. Khokhar, and A. Ghafoor. A framework for semantic modeling of video data for content-based indexing and retrieval. *ACM-Springer Multimedia Systems Journal*, 7(5):409–423, 1999.
6. M.E. Dönderler, E. Şaykol, U. Arslan, Ö. Ulusoy, and U. Güdükbay. BilVideo: A video database management system. *submitted journal paper*, 2002.
7. M.E. Dönderler, Ö. Ulusoy, and U. Güdükbay. A rule-based video database system architecture. *Information Sciences*, 143(1-4):13–45, June 2002.
8. M.S. Hacid, C. Decleir, and J. Kouloumdjian. A database approach for modeling and querying video data. *IEEE Transactions on Knowledge and Data Engineering*, 12(5):729–750, 2000.
9. J.Z. Li, M.T. Özsu, and D. Szafron. Modeling and video spatial relationships in an object database management system. In *Proc. of the International Workshop on Multimedia DBMSs, Blue Mountain Lake, NY*, pages 124–133, 1996.
10. Y. Rui, T.S. Huang, and S. Mehrota. Constructing table-of-content for videos. *ACM-Springer Multimedia Systems Journal*, 7(5):359–368, 1999.
11. E. Şaykol, U. Güdükbay, and Ö. Ulusoy. A semi-automatic object extraction tool for querying in multimedia databases. In S. Adali and S. Tripathi, editors, *7th Workshop on Multimedia Information Systems MIS'01, Capri, Italy*, pages 11–20, November 2001.