Towards a Two – Layered Video Metadata Model

Chrisa Tsinaraki, Stratos Papadomanolakis, Stavros Christodoulakis
Laboratory of Distributed Multimedia Information Systems and Applications
Technical University of Crete (MUSIC/TUC)
{chrisa, stratos, stavros}@ced.tuc.gr

Abstract

In this paper, we propose a model for video metadata that supports video retrieval based on video content, video structure and/or video attributes. Our model supports video retrieval based on the relationships among videos and between videos and real world objects. This model is also appropriate for providing personalization and recommendation functionality in video based services, while it takes into account events covered by more than one cameras. Our model is two-layered: In the first layer, a set of core classes appropriate for supporting any video type (e.g. news, movies, football matches, etc.) is defined. In the second layer, we define a set of classes, specific for each video type, that permit a more complete description of the videos of that type. We decided to implement our model using the functionality provided by MPEG-7.

1. Introduction

Information Retrieval is a key issue in everyday life, in the sense that everybody needs to discover the information he needs on a daily basis in order to accomplish the tasks contained in his routine. The Web has made this task much easier, as it made possible end-user access to information of interest. The advances in Web technology together with the mature relational database technology made possible the development of *Digital Libraries* on the Web. In this context, Digital Libraries are large collections of information that may be either of general interest or of interest to the members of a specific research community. Independently of the category of their contents (e.g. sports, politics, science etc.), the favored medium in the Digital Libraries available today is text. Even if multimedia are provided together with the textual data, multimedia objects are not usually of major importance. On the other hand, people prefer information in video format. Consequently, a number of video-based services have evolved (e.g. Digital Video Libraries, Video on Demand, Personalized TV etc), although the frameworks needed for their efficient support have not been developed yet.

One of the main drawbacks is the support of efficient video retrieval functionality. A solution for this problem may be given through the development of large *Digital Video Libraries*, which will provide video retrieval functionality comparable to that provided for textual data in the Digital Libraries available today. Although the general principles of information retrieval methodologies developed for text are still applicable in the digital video environment, satisfactory information retrieval techniques specific to digital video have still to be developed.

Information Retrieval services in general and personalization and recommendation in particular are based on *Metadata*, which are "data about data". Video metadata are data used for the description of video data, including the attributes and the structure of videos, video content and relationships that exist within a video, among videos and between videos and real world objects. The more complete a video metadata model is, the better the quality of the video retrieval services provided by the system based on the metadata model. Thus, in order to provide adequate information retrieval capabilities in video based services, an appropriate model for video metadata must be provided.

In this paper, we propose a model for video[1] metadata that supports video retrieval based on video content, video structure and/or video attributes. Our model supports video retrieval based on the relationships among videos and between videos and real world objects. This model is also appropriate for providing personalization and recommendation functionality in video based services. Our work is also inspired from the requirements identified and the ideas expressed by the TV-Anytime forum [12]. Additionally, we take into account events covered by more than one cameras.

Our model is two-layered: In the first layer, a set of core classes appropriate for supporting any video type (e.g. news, movies, football matches, etc.) is defined. In the second layer, we define a set of classes, specific for

[1] We make no assumption about the existence of a separate audio track. Here the word video implies unified video and audio content. However, in order to obtain the precise MPEG-7 semantics, one should replace every occurrence of the word Video with the word Audiovisual.

each video type, that permit a more complete description of the videos of that type.

We are implementing our model using the functionality provided by the MPEG-7 [11] standard (currently under development), a standard used for the description of video metadata. The MPEG-7 standard defines a set of *Description Schemes (DSs),* essentially complex data types that will be used to describe audiovisual content. The language used in the definition of the standard is the MPEG-7 Description Definition Language (DDL). Our model is implemented using MPEG-7 DDL and will be integrated to the metadata management system of the UP-TV[2] project. Among the objectives of the UP-TV project is to provide functionalities for content-based selection of videos or parts of videos to record from a broadcast or download from a server.

In the rest of the paper, we present our core metadata model, we propose the usage of application specific extensions for the provision of more adequate support for certain classes of applications and we discuss our conclusions and the future directions of our research.

2. Video Metadata Model

In this section, we describe our metadata model and present its usage through a set of examples. A key aspect for the definition of a video metadata model is the imposed video structure. Video data are often represented either as a set of still images that contain salient objects [9] or as clips that have specific spatial (e.g. color, position etc.) or temporal (e.g. motion) features or are related to semantic objects [10] [3] [2]. More sophisticated approaches are either a hierarchical representation of video objects [1] [13] [8] [5] based on their structure, or an event-based approach that represents a video object as a set of (non-contiguous, even overlapping) video segments called *strata* or *temporal cohesions* [7] that correspond to individual events. An interesting approach is taken in [4], where a hierarchical video structure based on timelines is defined and appropriate annotations are attached to the different levels of the video structure.

The metadata model we propose combines ideas from both the hierarchical and the event-based approach in order to provide a powerful set of information retrieval capabilities that cover the needs of different user communities and support the provision of personalization and recommendation functionality. In addition, the model takes into account the relationships between video objects and real world objects [6] [1] and events covered by more than one cameras.

Our model is structured in two layers: It provides a set of core classes capable of providing a basic level of support for every video type (core model), and a set of application-specific classes for the support of additional functionality for well-studied video types (application specific models). In this paper, we focus on the description of the core classes, while the need for application-specific classes is presented through a set of examples.

The imposed structure of a video, according to our core model, is shown in

Figure 1: A video is represented as an instance of the (Video) *Program* class and is comprised of a set of *Stories*. Each story is a logical section of the video object (e.g. a half-time of a football match, a news reportage in news etc.) and is further divided in a set of *Scenes*. A scene represents an event (e.g. a part of the news where the newscaster is talking or a penalty in a football match) and may be either *Composite* or *Simple*: a simple scene contains a simple event and is comprised of video *Shots*, while a composite one contains a composite event and is comprised of other (simple or composite) scenes. Shots are sets of "similar" consequent frames that are usually recognized using automatic segmentation techniques.

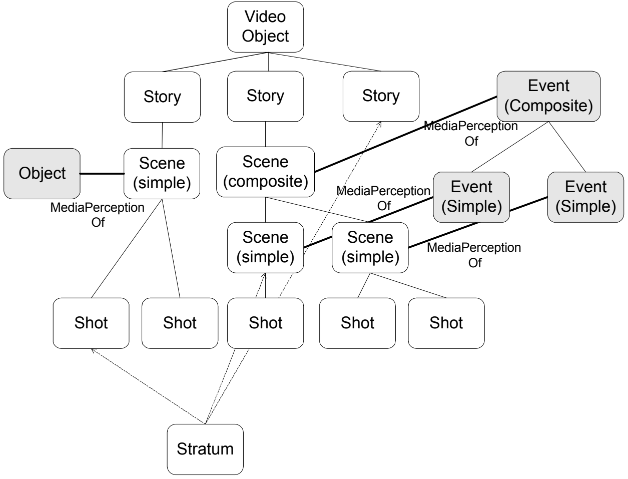

Figure 1: Video Object Structure

In addition to this hierarchical structure, we define *Strata* as non-contiguous video segments where certain events take place (e.g. the goals in a football match). A *Stratum* may be comprised of video segments belonging to one video or to several videos (e.g. all the goals scored by a certain player in the last year). The later is appropriate when a digital library of videos is maintained in a server, and a new video broadcast is cross-linked with archival video information to enhance the interactivity and openness of the user environment.

A video segment (shot, scene, story, video object or stratum) may be related to both *salient objects* and *events*

[2] UP-TV (Ubiquitous & Personalized TV Services) is a European R&D project, in which MUSIC takes part and is responsible for its metadata management system.

(through the MPEG-7 defined relationships *MediaPerceptionOf* and *HasMediaPerceptionOf*). A salient object represents an important object that appears in a video segment, while an event represents an event that takes place in a video segment. The events and the salient objects together with the *MediaPerceptionOf* and *HasMediaPerceptionOf* relationship types comprise the part of our core model that relates video objects with real world objects.

As an event may be covered by more than one cameras, all the video segments in our model are in fact video segment collections. Every collection is comprised of a set of video segments each of which is characterized by the camera used for recording it. For every video segment in our model a "default" camera is selected and the corresponding segment is the one sent to the users if there is no other preference set by them. This can be overcome by the user preferences: If a user prefers one of the other cameras, the corresponding segment is sent to him as an alternative to the default one.

For example let three cameras cover a scene where a goal occurs: one on the centre of the stadium, one in front of the goalposts and one behind them. The scene is the collection of the video segments recorded from each of the cameras and each of them may be used as an alternative of the other ones. Let the camera in front of the goalposts be the default camera. Then, the default segment sent to the users is the one recorded by this camera. If a user has denoted in his preference profile that he prefers the camera behind the goalposts he receives in his TV-set the corresponding video segment instead of the default one.

A high level description of the model using a UML class diagram and omitting the attributes and functions from each class description is given in Figure 2. As is shown in Figure 2, the classes that represent contiguous video segments and define partitions of video programs (*Story*, *ComplexScene*, *SimpleScene* and *Shot*) are subclasses of the *ContiguousVideoSegment* class. The *ContiguousVideoSegment* class and the *Stratum* class are subclasses of the *VideoSegment* class, the MPEG-7 class that represents video segments. The relationship between a containing contiguous video segment and a contained one (e.g. a story and a scene) is the *Partition* relationship. Semantic information is represented by the *Object* and the *Event* classes (both subclasses of the MPEG-7 *SemanticBase* class) and their relationships (*MediaPerceptionOf* and *HasMediaPerceptionOf*) with video segments.

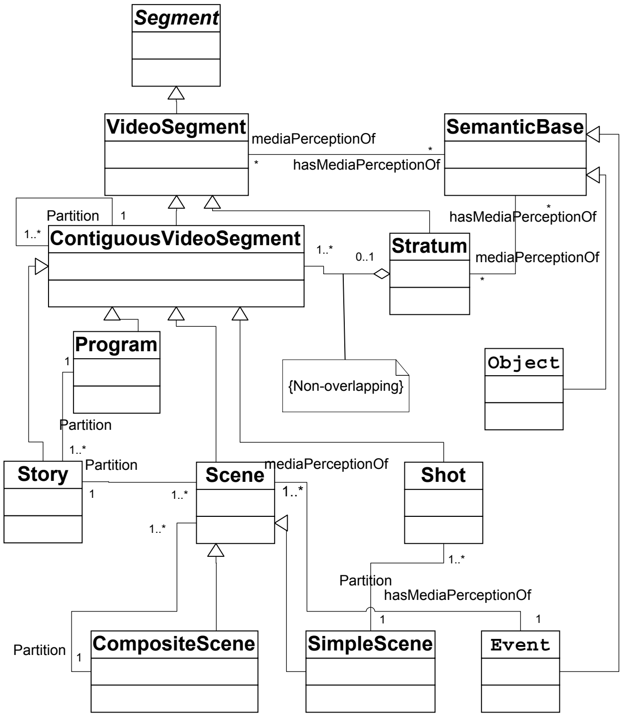

Figure 2: UML Class Diagram of the Video Metadata Model

As far as it concerns strata, we decided to define a set of Stratum subclasses, based on the work carried out in the TV-Anytime forum. The resulting class hierarchy is roughly depicted in the UML class diagram of Figure 3.

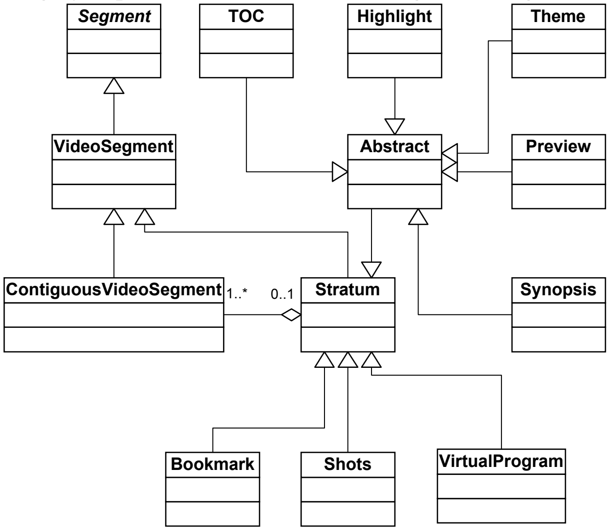

Figure 3: The *Stratum* class hierarchy

In Figure 3, we show how we extend our stratification approach in order to support the *Segment Group* functionality defined in the TV-Anytime forum Segmentation Metadata Specification. Thus, we derive from the *Stratum* class the following subclasses.

- *Bookmark*, which represents a stratum defined by a set of access points within a specific video program.

- *Shots*, which represents a stratum containing a collection of shots.
- *VirtualProgram*, which represents a stratum consisting of video segments taken from different video programs.
- *Abstract*, which represents a stratum that serves as an abstract for the content of a video program or a part of a video program. Based on the TV-Anytime forum requirements, we defined the following subclasses of *Abstract:*
  - *Table of Contents*, which represents a stratum consisting of segments that define a table of contents for the content of a video program.
  - *Highlights*, which represents a stratum consisting of segments that contain only the highlights of the content of a video program.
  - *Theme*, which represents a stratum consisting of segments that have a common theme. For example, a theme for a specific news item can be defined as a stratum containing all the segments of a news broadcast that refer that news item.
  - *Preview*, which represents a stratum consisting of a selection of segments that provide a preview of the content of a video program (for example for promotional purposes).
  - *Synopsis*, which represents a stratum containing a collection of segments that provide a summary of the content of a video program.

Strata as well as the subclasses of Stratum defined above are going to be implemented as the *Highlights* part of the *Summaries* provided by MPEG-7 for the definition of video program abstractions.

Each object belonging to the above-described hierarchies will be separately annotated with a rich set of attributes, comprising three standard MPEG-7 DSs. These are the *MediaInformationDS* (physical format, physical location, etc), the *CreationInformationDS* (title, actors, creation location, classification, etc) and the *UsageInformationDS* (distributor, parameters of the broadcast, financial data, etc).

The above model will be translated into MPEG-7 DSs. In developing our model, we remained within the MPEG-7 conceptual framework. Our class hierarchies are based on core MPEG-7 concepts, like the Segment, and it is therefore guaranteed that our DSs will fit into the existing DS structure. Specifically, we maintain full compatibility by using the DDL's facilities for derivation (extension or restriction) to derive our new DSs from the existing ones.

3. Application Specific Model Extensions

After the description of our core metadata model in the previous section, consider here a video library application environment where the model may be used. The video library contains videos of sport events. The following scenarios are supported by our model:

- A sports reporter is working on the reportage of a football match and wants to present that part of the match video where the goals are shown. This can be defined, as a stratum comprised of video scenes where goals take place and each of the scenes is related to a "goal" event.
- Another reporter works on a reportage for a football player, e.g. Pele, and wants to present the goals he has scored during his career. The video containing them can be defined as a stratum comprised of video scenes where Pele's goals take place and each of the scenes is related to a "goal" event and to the "Pele" object.
- An end-user is watching a film while a football match takes place, but he would like the film to be suspended whenever a goal is scored and the goal to be played back. If he denotes it in his preference profile, whenever a "goal" event occurs the appropriate video segment (usually a –composite or simple– scene) is sent to his TV-set for playback.

  If the user has denoted in his preference profile that he prefers the recordings of cameras placed in certain points (e.g. in front of the goalposts), if such a camera covers the match the video segment recorded by that camera is sent to him instead of the default one.
- Another end-user is watching a film while a football match takes place, but he would like to see the most important events of the match when the film finishes. If he has denoted which events are important for him (e.g. goals, penalties, corners etc.) in his preference profile, a stratum containing them is defined for him and is sent to his TV-set for playback after the film ends.

  If the user has also denoted that he would like that the video sent to him shouldn't take more than 10 minutes and which events are more important for him, some of the less important events wouldn't be included in the stratum sent to him. Thus, if the user has denoted that goals are more important for him than penalties, which in turn are more important than corners and goals and penalties take 10 minutes, no corners are included in the stratum sent to him.

The above examples show that the video metadata model we have presented up to now can support personalization and recommendation in video-based services. The personalization and recommendation functionality is provided independently of the video type, but the expressiveness of the metadata model will be greatly enhanced if specializations of the core entities are defined for each class of videos. Therefore, our model allows for application specific extensions of the core model.

To give an intuition of the idea, consider the application specific extensions for football matches. The following specializations could be included:

- The specializations of the salient objects that are the ball and the match actors (players, referee, coaches etc.),
- The specializations of the events that are "goal", "penalty", "foul", "corner" etc.
- The specializations of the contiguous video segments: The specializations of the *Story* class correspond to the half-times and the specializations of the *Scene* class correspond to the scenes where certain events take place (e.g. foul scenes).
- In addition to the "general purpose" strata that are defined manually for each video or for the coverage of the needs of specific users, predefined subclasses of strata for specific events (e.g. goals) that are important for all users can be defined as a specialization.

The above highlights of a specific application extension is used to illustrate the full scope of our two-layered video metadata model: The first layer is a core model applicable to all video types, while the second layer is a set of extensions of the core model, tuned and adapted to the needs of specific applications and video types. Thus, the core model can be used as it is during the system startup, and when specific applications and video types are studied, it is extended in order to provide added, application-specific functionality.

4. Conclusions – Future Work

In this paper, we presented a video metadata model that supports advanced information retrieval capabilities, including personalization and recommendation. It also covers the requirements posed by the TV-Anytime forum and supports the coverage of events using multiple cameras. The model will be implemented using the functionality provided by MPEG-7.

Our model is structured in two-layered and we focus here on the first layer, where a set of core classes is defined. The classes defined in the first layer may be used for the description of any video, independently of the video type it belongs. An extension of our core model allows the definition of second-layer, application-specific classes that may be used for the description of well-studied video types. We are currently working on the definition of the set of extension classes needed for the adequate description of two video application environments: Football matches and news.

Another direction is the integration of our model with a working system, in order to study its usability in real-world situations. This will take place in the system that will be developed in the context of the UP-TV project, where our model will be the basis for the UP-TV metadata management subsystem.

5. References

[1] Analyti and S. Christodoulakis, *Multimedia Object Modeling and Content-Based Querying*, Proceedings of Advanced Course – Multimedia databases in Perspective, Netherlands 1995.

[2] Al-Khatib Q., Day F., Ghafoor A., Berra B., *Semantic Modeling and Knowledge Representation in Multimedia Databases*, IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 1, January/February 1999

[3] Dağtas, Al-Khatib W., Ghafoor A., Kashyap R. L., *Models for Motion-Based Indexing and Retrieval*, IEEE Transactions on Image Processing, Vol. 9, No. 1, January 2000

[4] Dumas M., Lozano R., Fauvet M.-C., Martin H., Scholl P.-C., *Orthogonally modeling video structuration and annotation: exploiting the concept granularity*. In Proc. of the AAAI'2000 Workshop on Spatial and Temporal granularities, 2000

[5] Günsel B., Ferman A. M., Tekalp A. M., *Video Indexing through Integration of Syntactic and Semantic Features*, Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision, Souasota, Florida, December 2-4, 1996

[6] Grosky W., Managing Multimedia Information in Database Systems, Communications of the ACM, Vol. 40, No. 12, December 1997

[7] Hacid M-S., Decleir C., Kouloumdjian J., *A Database Approach for Modeling and Querying Video Data*, IEEE Transactions on Knowledge and Data Engineering, Vol. 12, No. 5, September/October 2000

[8] Kyriakaki G., *MPEG Information Management Services for Audiovisual Applications*, Master Thesis, Technical University of Crete, March 2000

[9] Li J. Z., Özsu M. T., *STARS: A Spatial Attributes Retrieval System for Images and Videos*, Proceedings of the 4th International Conference on Multimedia Modeling (MMM'97), Singapore, November 1997, pages 69-84

[10] Li J. Z., Özsu M. T., Szafron D., *Modeling of Moving Objects in a Video Database*, Proceedings of IEEE International Conference on Multimedia Computing and Systems, Ottawa, Canada, June 1997, pages 336-343

[11] MPEG Group, http://www.cselt.it/mgeg

[12] TV-Anytime Forum, http://www.tv-anytime.org

[13] Yeo B-L., Yeung M., *Retrieving and Visualizing Video*, Communications of the ACM, Vol. 40, No. 12, December 1997