A Video Metadata Model supporting Personalization & Recommendation in Video-based Services

Chrisa Tsinaraki, Stratos Papadomanolakis, Stavros Christodoulakis
Laboratory of Distributed Multimedia Information Systems and Applications
Technical University of Crete (MUSIC/TUC)
P.O. Box 133, GR, 73100 Chania, Greece

Abstract *In this paper, we propose an MPEG-7 based video metadata model that supports personalization and recommendation in video-based services. This is achieved through the provision of adequate video retrieval support that includes video retrieval based on video content, video structure and/or video attributes, as well as video retrieval based on the relationships within videos, among videos and between videos and real world objects.*

Our video metadata model is structured in two layers: In the first layer, we define a set of core classes appropriate for the description of any video type. In the second layer, we define additional sets of classes, one for the detailed description of each video type.

We focus here on the detailed description of the first layer of our model, while we stress, through a set of examples, the necessity of the second layer.

Keywords: Video Metadata Model, MPEG-7,
Personalization, Recommendation

1 Introduction

Information Retrieval is a key issue in everyday life, in the sense that everybody has to discover information on a daily basis in order to accomplish his routine tasks. The Web has made information retrieval much easier, as it made possible end-user access to information of interest. Although access to information of interest is easier, simple access isn't sufficient: Users have to discover the information before accessing it, while they must also be informed for information that is potentially interesting for them. Some systems offer additional filtering of the information they provide, while some of them provide notifications for new information items that become available to potentially interested users. The above scenario is a simplified version of personalization and recommendation facilities that are both based on information retrieval technology. Thus, personalization and recommendation are important features of the services provided by modern information systems. Such mechanisms have been extensively developed for services based on textual data (e.g. Digital Libraries, News on Demand etc.). Since video is a favored medium for information consumers, services based on digital video have recently received much attention (e.g. Digital Video Libraries, Video on Demand, Personalized TV). Describing video content is different and more demanding than describing text content. Although the general principles of information retrieval methodologies developed for text are still applicable in the digital video environment, satisfactory personalization and recommendation 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 personalization and recommendation functionality 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 also supports video retrieval based on the relationships among videos and between videos and real world

[1] When we refer to video, we make no assumption about the existence of a separate audio track. Thus, the word video used here 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.

objects. In addition, our model is appropriate for providing personalization and recommendation functionality in video based services, while it has been also inspired from the requirements identified and the ideas expressed by the TV-Anytime forum [12].

Our model is two-layered: In the first layer, we define a set of core classes appropriate for supporting any video type (e.g. news, movies, football matches, etc.). In the second layer, we extend our model through the definition of a set of classes specific for each video type, that permit a more complete description of the videos of that type.

Our model has been implemented on top of a relational database, using the functionality provided by the MPEG-7 [11] standard (currently under development), a standard used for the description of video metadata. MPEG-7 defines a set of *Description Schemes (DSs),* essentially complex data types, which will be used to describe audiovisual content. The language used in the definition of the standard is the *MPEG-7 Description Definition Language (DDL)*. XML Schema Language [14] [15] [16] has been selected by the MPEG consortium as the MPEG-7 DDL.

The implementation of our model will be integrated in the metadata management system of the *UP-TV (Ubiquitous & Personalized TV Services)* project, a European R&D project, in which MUSIC takes part and is responsible for developing the UP-TV metadata management system. 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 make a detailed presentation of our core metadata model while we stress, through a set of examples, the necessity of application specific extensions for the provision of more adequate support for certain classes of applications. We close the paper with a discussion on our conclusions and the future directions of our research.

2 Video Metadata Model

In this section, we focus on the description of our core metadata model and present its usage through a set of examples. The description of our core model starts with a brief discussion on the state of the art in video representation, as the imposed video structure is a key aspect for the definition of a video metadata model.

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 definition of our metadata model combines ideas from both the hierarchical and the event-based approach in order to provide a powerful set of 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].

Our model is structured in two layers: On the first layer, it defines a set of core classes that comprise our *core model*, capable of providing a basic level of support for every video type. On the second layer, sets of application-specific classes for the support of additional functionality for well-studied video types, called *application specific models*, are defined. In this paper, we focus on the description of the core classes, while we stress the need for application-specific classes 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 execution 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 (e.g. a set of frames during a penalty execution covered by a certain camera with the same camera parameters – position, zoom, focus etc.).

In addition to the above-defined hierarchical structure, we define *Strata* as non-contiguous video segments where certain events take place (e.g. all 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. Stratum also covers the requirements posed by the TV-Anytime forum for the definition of non-contiguous video segments called *Segment Groups*.

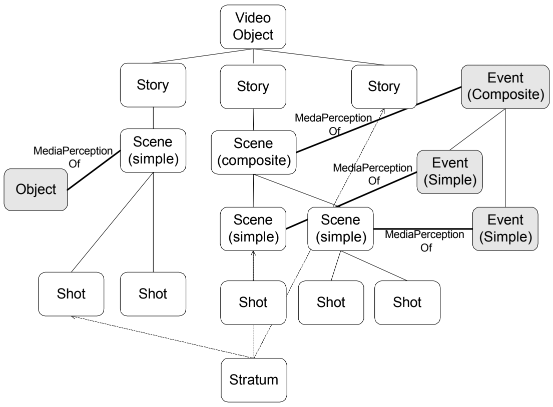

Figure 1: Video Object Structure

A video segment (shot, scene, story, video object or stratum) may be related to both *salient objects* and *events*, through the MPEG-7 defined relationships "*MediaPerceptionOf*" and "*HasMediaPerceptionOf*". A salient object represents an important object that appears in a video segment (e.g. the goalposts in a football match), while an event represents an event that takes place in a video segment (e.g. a goal scored in a football match). The events and the salient objects together with the relationship types "*MediaPerceptionOf*" and "*HasMediaPerceptionOf*" populate the part of our core model that relates video objects with real world objects.

A high level description of the model using a UML class diagram, where attributes and functions from each class description are omitted, 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, an abstract class that represents contiguous video segments. *ContiguousVideoSegment* and *Stratum* 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 and their relationships ("*MediaPerceptionOf*" and "*HasMediaPerceptionOf*") with video segments. Both Object and event are subclasses of the MPEG-7 *SemanticBase* class.

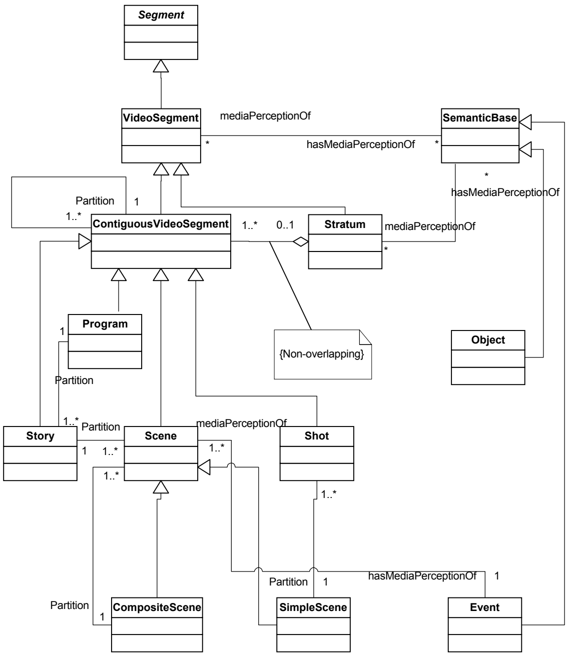

Figure 2: UML Class Diagram of the Video Metadata Model

Each object belonging to the above-described hierarchy 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 definition of the above DSs has been influenced by Dublin Core, a metadata standard for Internet resource description [18], while they are more detailed and fully cover the requirements for digital video description.

The above model has been 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 an integrated video service environment where our model may be used. The environment provides digital video library functionality as well as personalized broadcasting services in the context of sport events. Then, consider the following usage scenarios, having in mind that they are supported by our core metadata model:

- A sports reporter is working on a short reportage of a football match and wants to present that part of the match video where the goals are shown. The corresponding video segment 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.
- The same reporter is working on a more detailed reportage for the same football match, where goals, penalties, outs and corners are shown. The corresponding video segment can be defined, as a stratum comprised of video scenes where goals, penalties, outs and corners take place and each of the scenes is related to a "goal", a penalty, an out or a corner 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 belonging to different videos, where Pele's goals take place and each of the scenes is related to both a "goal" event and 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 has denoted 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.
- 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 their instances that take place in the match 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, according to his preferences, 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 for a certain match 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 video type. 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, the goalposts, the playground and the match actors (players, referee, coaches etc.). Thus, for each category of salient objects that appear in a football match, a corresponding subclass of the Object class should be defined (e.g. Ball, Goalpost, Playground, Player, Referee, Coach etc.).

- The specializations of the events that are “goal”, “penalty”, “foul”, “corner” etc. Thus, for each category of football match event, a corresponding subclass of the Event class should be defined (e.g. Goal, Penalty, Foul, Corner etc.).
- The specializations of the contiguous video segments: The specializations of the *Story* class for a football match correspond to the half-times and the specializations of the *Scene* class correspond to the scenes where certain events take place (e.g. goal scenes). Thus, a Half-Time subclass of the Story class should be defined for the representation of half-times. In addition, a subclass of the Scene class should be defined for each category of football match event in order to represent the scenes where instances of this event take place (e.g. GoalScene, PenaltyScene, FoulScene, CornerScene etc.).
- 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. strata that contain the goals scored in a football match) and their combinations (e.g. strata that contain the goals scored and the penalties executed in a football match) that are important for a large number of users can be defined as a specialization. Thus, for each category of football match event that is important for a large number of users or a favored combination of such events, a corresponding subclass of the Stratum class should be defined (e.g. GoalStratum, PenaltyStratum, FoulStratum, GoalPenaltyStratum etc.).

The above highlights of a specific application extension are 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 two-layer video metadata model, focusing 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.

The definition of application-specific models and the evolution of the requirements set by the TV-Anytime forum pointed out the need for providing functionality that was not taken into account originally in the definition of our core model. In order to enhance our model in terms of the functionality it provides, we have defined some extensions to our core model that:

- Support the coverage of events using multiple cameras [17].
- Fully support the TV-Anytime requirements as far as it concerns video segmentation, through the definition of a class hierarchy for different strata categories [17].

Another direction is the integration of our model with a working system, in order to study its usability in real-world situations. This is taking place in parallel, in the system that is being developed in the context of the UP-TV project, where our model will be the basis for the UP-TV metadata management subsystem. As far as it concerns implementation issues, our system is being implemented in a relational database (MySQL), while we use XSL style sheets for metadata presentation and for the definition of the actions that take place on video description insertion.

References

[1] A. 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

[14] Fallside D., XML Schema Part 0: Primer, http://www.w3.org/TR/xmlschema-0/, W3C Recommendation, May 2001

[15] Thompson H., Beech D., Malloney M., Mendelsohn N., XML Schema Part 1: Structures, W3C Recommendation, http://www.w3.org/TR/xmlschema-1/, May 2001

[16] Biron P., Malhotra A., XML Schema Part 2: Datatypes, W3C Recommendation, http://www.w3.org/TR/xmlschema-2/, May 2001

[17] Tsinaraki C., Papadomanolakis S., Christodoulakis S., Towards a two - layered Video Metadata Model, DLib Workshop 2001 (in conjunction with DEXA '01), September 2001 (to be appeared)

[18] S. Weibel, J. Kunze, C. Lagoze, M. Wolf, Dublin Core Metadata for Resource Discovery, September 1998, RFC 2413