SEMANTIC ADAPTATION OF MULTIMEDIA CONTENT

Michael Zufferey and Harald Kosch

Department of Information Technology, University of Klagenfurt, Austria
{michael.zufferey, harald.kosch}@itec.uni-klu.ac.at

ABSTRACT

The increasing diversity of devices and the heterogeneity of networks pose nowadays a challenge in the delivery and consumption of multimedia content. In this context, the Part 7 of the MPEG-21 standard formally named Digital Item Adaptation (DIA) targets the adaptation of multimedia content based on usage environment, such as network characteristics, terminal capabilities and user characteristics. But, MPEG-21 DIA does not take into account MPEG-7 semantics description tools, which provide means for a conceptual (semantic) description that is close to the human understanding of multimedia content. Therefore, to fill this gap, we propose an interactive and user-centric framework called *Semantic Adaptation Framework (SAF)*. The *SAF* provides facilities for the generation of all the required semantic metadata and enables an MPEG-21 adaptation engine to semantically adapt the multimedia content in order to provide the user with the best possible experience.

1. INTRODUCTION

The increasing diversity of multimedia enabled devices and the heterogeneity of networks pose several problems to multimedia content delivery and consumption. Typically, when a terminal accesses multimedia content to which it was not designed for, the user experience is rather poor. To enable the access to multimedia information by any terminal across any network is commonly referred in the literature as Universal Multimedia Access (UMA) [1][2]. In a UMA scenario and to enable an efficient adaptation of multimedia content, it is fundamental to have available descriptions of the parts that have to be matched and/or bridged:

- *Content description:* Information on the multimedia content characteristics such as coding format and parameters, resolution, bit-rate, color, spatial and temporal structure, concepts and objects (semantic).
- *Usage environment description:* Information on the user and usage context, such as user characteristics, terminal, network and natural environment.

These descriptions have to be generated, transmitted and processed by modules of the multimedia delivery and consumption chain in order to perform a set of adaptation operations to provide the user with the best possible experience. The new standard of MPEG, MPEG-21 [3][4] realizes UMA in many parts. The vision of MPEG-21 is to define a multimedia framework to enable transparent and augmented use of multimedia resources across a wide range of networks and devices used by different communities. To achieve this goal and to provide all the required functionalities (e.g., coding, multiplexing, synchronization, description, rights expression and management), MPEG-21 will make use of the relevant available standards such as MPEG-4 [5] for multimedia coding and representation and MPEG-7 [6][7] for content description and will develop new standards whenever required.

The Part 7 of the MPEG-21 called Digital Item Adaptation (DIA) deals with the adaptation of Digital Items[1] [8][9][10]. To enable Digital Item adaptation a (generic) Bitstream Syntax Description (gBSD) in XML format is used. This codec agnostic description of the bitstream enables a processor to adapt the bitstream without any further knowledge of the coding format. In a first stage, the adaptation engine is assumed to determine the optimal adaptation for the bitstream given the constraints as provided by the usage environment description and a steering description, which specifies the relationship between constraints, feasible adaptation operations satisfying these constraints, and possibly associated utilities (qualities) [8]. Based on that decision, the gBSD is consequently transformed and successively used to adapt the bitstream.

To adapt the multimedia content in order to best fit the user needs, MPEG-21 requires knowledge of the content itself. The better the content is known, the more efficiently it may be processed. In this context, the role played by MPEG-7 is essential. MPEG-21 relies on MPEG-7 for the description of the content and user interactions (user preferences and usage history), which are included in the MPEG-21 user characteristics. These characteristics allow filtering/searching (preferred topics, e.g., actors and movie genres or subjects) and browsing (content structures, e.g., key frames and video clips) of the multimedia content. But, no facilities are provided for the description of user preferences expressed in terms of semantic entities and their relations [11]. For example, it is not possible to describe user preferences like: "Arnold Schwarzenegger is fighting" and to adapt consequently the video.

The major focus of this paper is to provide an integration of MPEG-7 semantics description tools [11] into the MPEG-21 Multimedia Framework in order to provide semantic adaptation of the multimedia content and therefore enhance the user experience.

The remainder of this paper is organized as follows. Section 2 introduces first the semantic adaptation of multimedia content, then the *Semantic Adaptation Framework (SAF)* is presented and discussed. Section 3 describes the semantic adaptation process. Finally, conclusions and further work are given in Section 4.

[1] Digital Item is the exchange unit of multimedia data in MPEG-21. It is based on XML Schema and provides principally a container for metadata and media data.

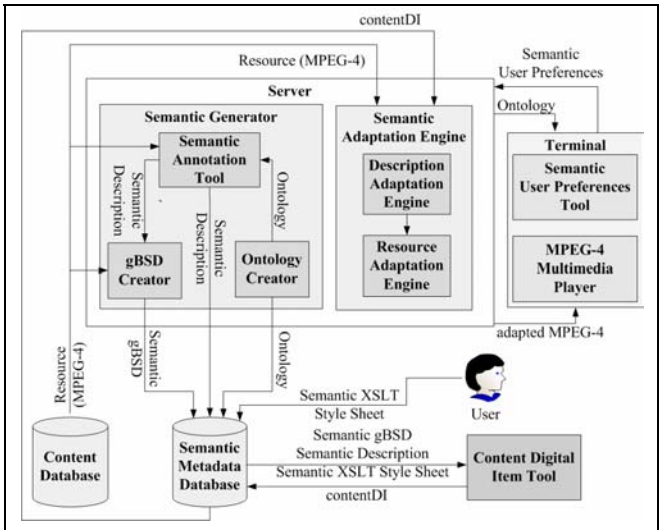

Figure 1: Architecture of the *SAF*

2. SEMANTIC ADAPTATION OF MULTIMEDIA CONTENT

Semantic adaptations involve the temporal and/or spatial reduction of specific multimedia content characteristics, e.g., temporal duration or number of regions of interest. The semantic adaptation may create a content summary by extracting temporal segments (e.g., key frames, audio clips) or spatial segments (e.g., regions of interest), which are relevant for the user. Semantic adaptations can be organized into two categories:

- *Adaptation by temporal summarization*: The content is temporally segmented into hierarchical structures (e.g., scene, clips, shots, frames), which are semantically annotated. This annotation relies on the user preferences, i.e., relevant temporal segments are selected to create the most adequate audiovisual summary for the concerned user.
- *Adaptation by spatial/scene summarization*: A scene has not only a temporal dimension, but also a spatial dimension which consists of various spatial segments, each of them with a semantic description [12]. The summarization process evaluates each spatial segment based on the user preferences. Unsuitable or unwished segments are removed or their quality/resolution is reduced.

The creation of summaries depends on the richness and accuracy of the MPEG-7 descriptions in terms of structure (temporal and spatial) and semantics of the multimedia content. In the following section we expose our solution framework for semantic adaptation.

2.1. Semantic Adaptation Framework

Our *Semantic Adaptation Framework (SAF)* provides facilities for the generation of semantic metadata required for the conceptual (semantic) adaptation of the multimedia content. Therefore, the content is not seen as an "assemblage" of low-level features (e.g., color, shape, texture), but in terms of semantic entities (e.g., events, concepts, objects, people, location, time) and their relations, which constitute what MPEG-7 names a "narrative world" [11].

The *SAF* specifies the following semantic metadata:

- *MPEG-7 compliant ontologies* for content annotation and context representation.
- *MPEG-7 semantic descriptions* of the resources.
- *Semantic generic bitstream syntax descriptions* (semantic gBSDs) of the resources, which consist of gBSDs [8] indexed with semantic metadata extracted from the MPEG-7 semantic descriptions.
- *Steering descriptions*, which define all the possible semantic adaptation for the resources.
- *Semantic user preferences*.
- *Content Digital Items*, which provide a flexible link between the semantic gBSDs and the steering descriptions.

The modular architecture of the *SAF* (Figure 1) allows an easy extensibility and scalability of both software modules and semantic metadata.

The *SAF* consists of the following modules:

- *Semantic Generator*: Provides means for the generation of the ontologies, for the semantic annotation of the resource and generation of the MPEG-7 semantic description. It also generates the semantic gBSD and the steering description.
- *Semantic Adaptation Engine*: Semantically adapts the resource based on the semantic user preferences and the MPEG-7 semantic description.
- *Content Digital Item Tool (CDITool)*: Builds a content Digital Item (contentDI) by aggregating the semantic gBSD, the steering description and other content related metadata.
- *Semantic User Preferences Tool*: Allows a user to select her/his preferred topics, content structures and semantic entities with their relations from the available ontologies.

A typical walkthrough can be summarized as follows. First, the *OntologyCreator* is used by a user to generate one or more ontologies specific to her/his domains of interest. Then she/he makes use of the functionalities provided by the *Semantic Annotation Tool (SAT)* to annotate the desired MPEG-4 resource based on the ontologies and to generate an MPEG-7 semantic description. The *gBSDCreator* is successively used to parse the resource, generate the semantic gBSD and index it with the semantic metadata extracted from the MPEG-7 semantic description. The user may also create a semantic XSLT style sheet which enables an adaptation engine to transform the semantic gBSD based on the adaptation decision. Finally, the *CDITool* aggregates the semantic gBSD, the MPEG-7 semantic description and the semantic XSLT style sheet into a contentDI.

To collect semantic user preferences (context), a connection is established between the server and the terminal. The ontologies are sent to the terminal and used by the *Semantic User Preferences Tool* to allow a user to select her/his preferred topics, content structures and semantic entities with their relations. The generated semantic user preferences are then sent back to the server which stores them into the *Semantic Metadata Database*.

When the user requests to watch a video, a connection is established between the terminal and the server. The *Semantic Adaptation Engine* retrieves from the Semantic Metadata Database the semantic user preferences and the contentDI related to the requested video. Based on the semantic metadata stored in the contentDI and the semantic user preferences, the *Semantic Adaptation Engine* semantically adapts the video and delivers it to the terminal. Finally, the user can watch her/his desired adapted video using the *MPEG-4 Multimedia Player*.

```
<Mpeg7>
 <Description
  xsi:type="ClassificationSchemeDescriptionType">
 <ClassificationScheme uri="urn:saf:cs:UserInterestCS">
  …
  <Term termID="1" id=”O1”>
   <Name xml:lang="en">Animal reign</Name>
   <Term termID="1.1" id=”O1.1”>
    <Name xml:lang="en">Invertebrates</Name>
    <Term termID="1.1.1" id=”O1.1.1”>
     <Name xml:lang="en">Corals</Name></Term>
    …
   </Term></Term>
 </ClassificationScheme></Description></Mpeg7>
```

Figure 2: Example of ontology which describes user interests (partial)

2.2. Ontologies for Content Annotation and Context Representation

The semantic adaptation of multimedia content requires knowledge about the content (semantic description) and the context (semantic user preferences). The preferences of the concerned user are matched to the content description in order to produce an adaptation decision which is used to adapt the content.

The *SAF* defines specific ontologies compliant to MPEG-7 for the standardized representation of the following semantic metadata:

- *Topics:* Preferences of a user for specific content thematic (e.g., preferred movie or actor) and subjects of interest (e.g., animals, archaeology and geology).
- *Structures:* Preferences of a user for a particular content digest such as key frames, video clips, audio clips, etc.
- *Semantic entities and relations*, e.g., objects, agent objects, events, concepts, relations between semantic entities.

Topics and structures metadata are described by using MPEG-7 Classification Schemes (CSs) [11], which enable a structured description of a set of terms and their relations specific to a particular domain. Instead, semantic entities and their relations rely on MPEG-7 semantics description tools. These ontologies provide an efficient and flexible bridge between content annotation and user context description.

Figure 2 depicts an example of an ontology used to describe user interests. The attribute *id* is used to specify the type of the *Term* node (e.g., O=ObjectType) and therefore to simplify the content annotation process.

The ontologies are used by the *SAF* for the annotation of multimedia content and the generation of semantic descriptions based on the MPEG-7 semantics description tools. Figure 3 describes an example of semantic description based on the ontology depicted in Figure 2.

These descriptions are semantic-centric which is explained in the following. The resource is not segmented into scenes, shots and frames and then semantically annotated. Instead, semantic entities and their relations are first detected and then their temporal position and duration is annotated. The temporal annotation is used to map the semantic entities and their relations to the corresponding *gBSDUnits* of the semantic gBSD.

```
<Mpeg7><Description xsi:type="SemanticDescriptionType">
 <Semantics><Label>
  <Name>Video on Marine Life</Name></Label>
 ...
  <SemanticBase xsi:type="ObjectType">
   <Label href="urn:saf:cs:UserInterestCS:1.1.1">
    <Name>Corals</Name></Label><MediaOccurrence>
   …
    <VideoSegment><MediaTime>
     <MediaTimePoint>T00:00:00:00F25</MediaTimePoint>
     <MediaDuration>PT3S1N25F</MediaDuration>
    </MediaTime></VideoSegment>
   …
  </MediaOccurrence></SemanticBase>
 ...
</Semantics></Description></Mpeg7>
```

Figure 3: Fragment of MPEG-7 semantic description

The ontologies are also used by the *Semantic User Preferences Tool* to enable a user to build/edit her/his semantic preferences (context). The generated semantic user preferences are successively used during the adaptation process.

3. ADAPTATION PROCESS

The *Semantic Adaptation Engine* is responsible for the semantic adaptation of the resource based on the user preferences. The semantic adaptation process can be summarized as follows. First the *Description Adaptation Engine* computes an adaptation decision based on the steering description and the semantic user preferences. The adaptation decision consists of a parameterized XSLT style sheet, which is used to transform the semantic gBSD. The transformed semantic gBSD is successively used by the *Resource Adaptation Engine* to adapt the resource.

The semantic generic bitstream syntax description (semantic gBSD) provides a high-level structure description of the resource. Figure 4 depicts a fragment of an MPEG-4 Video Elementary Stream (VES) semantic gBSD. In this example, the first two gBSDUnits describe the Video Object (VO) and the Video Object Layer (VOL) headers, which contain configuration information for the decoding of the VES [13]. Each successive *gBSDUnit* describes one Video Object Plane (VOP) [5][13] and the *syntacticalLabel* attribute specifies the type of the VOP. The *marker* attribute describes the semantic metadata extracted from the MPEG-7 semantic description depicted in Figure 3 (the length parameter specifies the length in bytes of the semantic segment), the composition time stamps (CTS) and the decoding time stamps (DTS) [5][13]. Knowledge of the CTS and DTS for each VOP is required for their re-synchronization, because semantic adaptation by removing unsuitable and/or undesired VES segments causes their de-synchronization. The de-synchronization of the CTS and DTS results for the user to watch a black screen for the duration of the removed segment.

Generally, the parsing of the MPEG-4 VES and the generation of the semantic gBSD pose no significant problems. Instead, the semantic indexing is a more complex process due to the structure of the MPEG-4 VES, which comprises several VOPs classified as different types, such as I-, P- and B-VOP among others [5][13]. As the semantic adaptation consists of removing unwished and unsuitable VES segments, it is essential

to index the semantic gBSD in such a way that removed segments do not affect the decoding process. Therefore, the *gBSDCreator* first performs a semantic gBSD structure analysis in order to identify the type and the pattern of the used VOPs. Then, it extracts the VES segments representing the semantic entities and their relations from the MPEG-7 semantic description and uses the temporal information stored in their *MediaTime* node to map them to the corresponding *gBSDUnits* of the semantic gBSD. The first *gBSDUnit* of each VES segments is indexed with the corresponding semantic metadata and with a length parameter that represents the length in bytes of the VES segment. If a corresponding *gBSDUnit* represents an unsuitable[2] VOP, the nearest suitable VOP is searched in either directions (backward and forward) and used.

```
<dia:DIA>
 …
 <dia:Description xsi:type="gBSDType" id="MarineLifegBSD"
  bs1:bitstreamURI="MarineLife.cmp">
  <gBSDUnit syntacticalLabel=":MV4:VO" start="0"
   length="4"/>
   <gBSDUnit syntacticalLabel=":MV4:VOL" start="4"
    length="22"/>
   <gBSDUnit syntacticalLabel=":MV4:I_VOP" start="26"
    length="2107" marker="CTS=1 DTS=0
    urn:saf:cs:UserInterestCS:1.1.1#length=72487"/>
   …
   <gBSDUnit syntacticalLabel=":MV4:B_VOP" start="72513"
    length="336" marker="CTS=75 DTS=75"/>
   …
 </dia:Description></dia:DIA>
```

Figure 4: Fragment of an MPEG-4 VES semantic gBSD

MPEG-21 proposes the Adaptation QoS (AQoS) to steer the adaptation process [8]. The AQoS is based essentially on quantitative properties of both terminal and network. Instead, semantic adaptation targets qualitative properties of the multimedia content, such as topics, content structures and semantic entities and their relations. Moreover, the AQoS takes as input a single value of a specific constraint (e.g., network bandwidth) and generates as output the adaptation decision for that constraint (e.g., B-VOPs dropping). To cope with semantic adaptation one has to use multiple input and output values for specific constraints (e.g., user interests). For example, a user is normally interested in more than one element of a specific topic (e.g., Arnold Schwarzenegger and Harrison Ford as preferred actors) and in more than one topics (e.g., actors and animals). Moreover, each input value can generate more than one output value (e.g., animals may result in fishes and dogs). In this case, the use of the MPEG-7 semantic description as alternative to the AQoS to steer the adaptation process is more appropriate and therefore used within the *SAF*.

The *SAF* also defines the semantic XSLT style sheet, i.e., a parameterized XSLT style sheet to perform an efficient transformation of the semantic gBSD and the contentDI which provides a flexible link between the MPEG-7 semantic description (steering description) and the semantic gBSD (a reference to the bitstream is also included).

[2] Unsuitable means if it is removed from the VES, it will prevent VOPs which depend on it to be decoded.

4. CONCLUSIONS AND FURTHER WORK

Our *Semantic Adaptation Framework (SAF)* is a first approach for integrating MPEG-7 semantic metadata (topics, content structures and semantic entities with their relations) into the MPEG-21 Multimedia Framework. The *SAF* is an interactive and user-centric framework, which provides new functionalities, such as extensible ontologies for the unified and MPEG-7 standardized representation of semantic metadata, fine-grained semantic indexing of the semantic gBSD and content Digital Item to link in a flexible and extensible way the semantic metadata needed for the adaptation process.

Further work will lead to the implementation of a learning algorithm for the semi-automatic annotation of semantic entities, the implementation of an optimized adaptation engine for semantic adaptation and to the integration of a streaming server/client for an efficient delivering of adapted multimedia content.

5. ACKNOWLEDGEMENT

This work is partially supported by the EC DANAE project (IST-1-507113) [14].

6. REFERENCES

[1] Special Issue on Universal Multimedia Access, *IEEE Signal Processing Magazine*, 20(2), March 2003.

[2] Special Issue on Universal Multimedia Adaptation, *Signal Processing: Image Communication*, 18(8), September 2003.

[3] ISO/IEC 21000-1:2004, Information technology – Multimedia framework (MPEG-21) – Part 1: Vision, Technology and Strategy, Munich, March 2004.

[4] I. Burnett, R. Van De Walle, K. Hill, J. Bormans, F. Pereira, "MPEG-21: goals and achievements", *IEEE Multimedia*, 10(4), pp. 60-70, October-December 2003.

[5] T.F. Ebrahimi, F. Pereira (eds.), *The MPEG-4 Book*, Prentice Hall PTR, New Jersey, July 2002.

[6] B.S. Manjunath, P. Salembier, T. Sikora (eds.), *Introduction to MPEG-7: Multimedia Content Description Language*, John Wiley & Sons, West Sussex (England), June 2002.

[7] H. Kosch, *Distributed Multimedia Database Technologies Supported by MPEG-7 and MPEG-21*, CRC Press LLC, Florida, November 2003.

[8] ISO/IEC 21000-7:2004, Information technology – Multimedia framework (MPEG-21) – Part 7: Digital Item Adaptation, 2004.

[9] A. Vetro, "MPEG-21 Digital Item Adaptation: Enabling Universal Multimedia Access", *IEEE Multimedia*, 11(1), pp. 84-87, January-March 2004.

[10] A. Vetro, C. Timmerer, "Digital Item Adaptation: Overview of Standardization and Research Activities", *To appear in: IEEE Transactions on Multimedia*, 2005.

[11] ISO/IEC 15938-5:2003, Information technology – Multimedia Content Description Interface – Part 5: Multimedia Description Schemes, May 2003.

[12] K. Nagao, Y. Shirai, K. Squire, "Semantic Annotation and transcoding: making web content more accessible", *IEEE Multimedia*, 8(22), pp. 69-81, April-June 2001.

[13] ISO/IEC 14496-2:2003, Information technology – Coding of audio-visual objects – Part 2: Visual, Pattaya, March 2003.

[14] DANAE Website: http://danae.rd.francetelecom.com