Context-Aware Semantic Adaptation of Multimedia Presentations

Mariam Kimiaei Asadi
Ecole Nationale Supérieure de Télécom.
46 rue Barrault, 75013 Paris, France
Mariam.Kimiaei@enst.fr

and

Jean-Claude Dufourd
Streamezzo
83, Bd du Montparnasse, 75006 Paris
Jean-Claude.Dufourd@ Streamezzo.com

Abstract

In this paper, we present our work on context-aware semantic adaptation of multimedia structured documents. We introduce the semantic annotation of a multimedia scene, providing semantic information on each media object of the scene, as well as on the dependencies between all the media objects. The proposed description tools are based on the framework of MPEG-21. We use these semantic annotations in order to perform a semantic adaptation on multimedia presentations. We aim to show that, in order to preserve the consistency and meaningfulness of the adapted multimedia scene, the adaptation process needs to have access to the semantic information of the presentation.

1. Introduction

Over the past several years, the development of information technology and growth of multimedia popularity as well as user demands have led to the creation of a vast variety of multimedia content and devices. Delivery of such diverse multimedia content to different types of user devices and environments is one of the major challenges of a multimedia delivery chain. Content delivery chains need to have enough information on the context of the usage (network, device, and user preferences) of the multimedia content in order to be able to provide the end user with the optimum form of the content.

A knowledge-based and semantic multimedia adaptation infrastructure is then needed to satisfy these requirements. Such an infrastructure should propose methods to express context constraints, as well as, content structural and semantic information. MPEG (*Moving Picture Experts Group*) and W3C (*World Wide Web Consortium*) have provided recommendations and standards, which support and define frameworks for a multimedia content adaptation system. These standards, however, do not provide complete support for semantic adaptation. MPEG-21, on which we have based our work, pays special attention to the support of resource (single media) adaptation based on context constraints. Requirements for semantic adaptation of composed multimedia presentations are not considered in MPEG-21. In this paper, we present our work on a semantic and context-aware multimedia adaptation framework based on MPEG-21 [1]. The process of adaptation of each resource and of the whole multimedia presentation, takes into account the constraints of the context as well as the semantic metadata of the multimedia content (both single media objects and the whole presentation). Semantic adaptation of multimedia presentations, addresses the adaptation of multimedia structured documents based on temporal, spatial and semantic relationships between the media objects.

In this paper, we first give a short introduction to MPEG-21. We next give a short summary on the state of the art and the adopted approaches in the area of semantic adaptation of multimedia *scenes*. Next, the principles elements of our approach, Multimedia Scene Semantic Adaptation (MSSA), and the strategy and methods behind it are described. Finally, the last section offers conclusions on the presented approach as well as our perspective on this work.

2. MPEG-21

MPEG-21 is an ISO standard from the MPEG family that identifies and defines the key elements needed to support a multimedia delivery chain, the relationships between and the operations supported by them. The basic notion of MPEG-21 is that of the “Digital Item” (DI). A “Digital Item” is a multimedia content and its related metadata. As described in MPEG-21 a “Digital Item” is the digital representation of “a work”, and as such, it is the thing that is acted upon (managed, described, exchanged, collected, etc.).

MPEG-21 consists of several parts. The parts on which we have based our work are Digital Item Declaration (DID), Digital Item Adaptation (DIA).

3. Why semantic adaptation?

A multimedia *scene* (this vocabulary has been adopted from MPEG-4 [2]) is a synchronized multimedia presentation that integrates multiple static, or continuous media. It also specifies how they should be combined together and, based on spatial and temporal factors, be presented to the user. There exist several languages for describing multimedia scenes. MPEG has defined an XML-based description language for MPEG-4 scenes, called XMT. BIFS (*BInary Format for Scenes*) is the binary format of this scene description language [2]. SMIL (*Synchronized Multimedia Integration Language*) [3], a W3C recommendation, is an XML-based scene description language with strong temporal functionalities.

When adapting a multimedia presentation, in order to preserve the consistency and meaningfulness of the adapted scene, the adaptation process needs to have access to the semantic information of the presentation.

For instance, consider one image media and its text caption within a multimedia presentation. If, throughout the process of adaptation, the image is eliminated because of a bandwidth limitation, or the lack of image support by the terminal, the adaptation engine should also remove the caption of the image. This is not feasible without having semantic information on the scene. Another simple example is a multimedia document with two images and two texts, each text giving explanation on one of the two images. Let’s assume that the display size of the user device is too small for the whole scene, even after maximum downscaling of the images. A fragmentation of the scene then becomes necessary. In this case, in order to keep the related image and text together in the same scene fragmentation, and to temporally sort the fragmentations in the correct order, the adaptation engine needs some semantic information on the scene.

As seen in these examples, a complete content adaptation sometimes requires a good understanding of the original document. If the adaptation process fails to analyze semantic structure of a document, then the adaptation result may not be accurate and may cause user misunderstanding or non-comprehension.

4. Related work

While numerous different approaches have been adopted in the area of resource (single media) adaptation, less work has been done on the semantic adaptation of multimedia scenes. Mohan et al. present solutions on adaptation of multimedia presentations, based on some limited semantic information, mainly on the purpose of image media objects. This information is not explicitly given and is obtained from the original image object [4]. F. Rousseau et al., also propose solutions for the adaptation of multimedia presentations that remain incomplete from the semantic point of view [5]. J. Euzenat et al., present solutions for adaptation of multimedia documents only along their temporal dimension [6]. In the area of *Semantic Web* (http://www.w3.org/2001/sw/) several research activities have been done on the ontology-based semantic description of Web documents based on RDF (http://www.w3.org/RDF/). Nagao propose external semantic annotation in order to make Web documents adaptable [7]. The proposed semantic information remains incomplete concerning dependencies between media objects. Hori proposes semantic annotation and adaptation of HTML documents [8].

5. Multimedia Scene Semantic Adaptation (MSSA) framework

Our semantic adaptation system for multimedia presentations is built upon five principle elements:

5.1. Content description

Explicit description of the content is very importance to a multimedia content adaptation framework. A content adaptation engine needs to have an exact and complete description of the original content. Although it is true that some of the characteristics of the content could sometimes be directly extracted from the content itself, such as its modality or format, there are, nevertheless, some characteristics of the content that need to be given explicitly, such as some semantic information (e.g. semantic key frames of a video content, as opposed to “encoding” key frames), encoding parameters or other parameters, e.g. the maximum spatial downscaling of each visual media, with which it is still logically visible (we call this *maxRRF:* maximum Resolution Reduction Factor).

Among W3C recommendations, RDF is designed for content description. Within MPEG standards, MPEG-7 provides complete description tools for content description (http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm). We have used MPEG-7 descriptors for this purpose. For each media object of the scene, descriptors are wrapped up in a DID

Statement element and attached to a *Resource* element containing a reference to the concerned resource (media object). We call CDI (Content Digital Item) the DID document which contains the references to the content and its description.

5.2. Context description

The description of the context plays also an extremely important role in a multimedia content adaptation framework. A knowledge-based multimedia content adaptation framework needs to have exact information on the context (network, device, user, etc.) of the usage of the multimedia content in order to be able to provide the end user with the optimum form of the content. W3C defines CC/PP (http://www.w3.org/Mobile/CCPP/), based on RDF for context description, while MPEG defines MPEG-21 DIA for this means. We use MPEG-21 DIA, for the description of context. We call XDI (conteXt Digital Item) the DID document describing the context of the usage.

5.3. Semantic Information Description (SID)

Our semantic adaptation system requires an in-depth understanding of the document. Therefore, it needs human intervention. The semantic information of a multimedia scene could be either given by the author of the document or by any other editor. We have defined XML schemes as extensions to MPEG-21 DIA for the expression of semantic information of a multimedia scene. In the CDI, these descriptors are given in a *Statement* element attached to *Resource* element containing a reference to the concerned multimedia scene. The SID (Semantic Information Description) descriptors are used by the adaptation engine to decide on the type and nature of the adaptation(s) and then apply them to the scene. The information included in SID descriptors is categorized into three main parts: independent semantic information of each media object, semantic dependencies between media objects of the scene, and semantic preferences on scene fragmentations.

The first category describes, for each media object, its independent semantic information in the context of the scene, such as, importance and/or role (for example if it has a key role, it should not be, in any case, removed or degraded). The second category includes spatial dependencies (i.e., which media objects should be kept close together), absolute semantic dependencies (i.e., which media object is, or could be, a precondition, or redundant to another media object), and temporal dependencies (i.e., synchronization information between media objects). The third category describes preferences (i.e., priorities) of the spatial and temporal fragmentation.

5.4. Scene description

In our approach, we use SMIL for describing scenes. Note: our methodology is independent from the choice of SMIL. This could be done for other multimedia description languages. The rationale behind this choice is that SMIL is a high level scene description language, therefore, performing adaptations on a SMIL scene is easier compared to, for example, performing adaptations on a XMT scene. We map the media objects of the SMIL scene to media objects in the DID instance.

5.5. Scene optimization

In this section we describe our *scene optimizer*. Figure 1 shows the architecture of this module. The inputs are: description of context (XDI), semantic information and *physical* description of content (CDI), media objects and the SMIL scene. The latter could also be given (referenced) through CDI.

We consider the following rules: Resource transmoding [9] is a pure modality conversion process and has no effect on the spatial size (resolution) of a visual media. Only low-importance and redundant resources can be removed. The display size of the target device is always smaller than the original layout of the scene. Despite having defined these rules, the algorithm of the *scene optimizer* is still quite complex. The reason is that we have to, simultaneously, optimize both the resource and the scene adaptations. Temporal synchronizations are also complicating factors.

The scene optimizer first verifies the modality support of the target device and then removes the resources of the non-supported modalities. It then attempts to replace these medias by other medias in other modalities using the content description (attached to resource in CDI). If the attempt proves unsuccessful, it simply removes them from the scene. In every step of the optimization, when a *key role* media is to be removed, adaptation is considered to be impossible and the optimizing process is cut; we call this an *impossible adaptation* case.

If even after maximum downscaling of the resources (using maxRRF from CDI), the target display size is still too small to show the whole downscaled scene, based on the information given in SID descriptors, groups of semantically related media objects are constructed. These groups are then sorted by their timing priorities (given in the SID descriptor). Starting with the object group of the highest timing priority, the overall original spatial size (resolution) of the group is then calculated for each group. In case this is smaller than the size of the target device, we produce a scene fragment containing objects of this group. And if not, using the *maxRRF* of each media object of this group, we calculate the minimum possible overall spatial size of this group. Then, if this is smaller that the target device display, we calculate the optimum transcoding (resizing) of the media objects (this optimum resizing is calculated based on each media original size and it's importance), so that the overall group resolution becomes equal or smaller than the target display. If the minimum possible resolution of the group is not smaller than the display size we drop off redundant medias or medias of low importance, or replace them by a hyperlink pointing to them, and we then recursively redo the calculation for this new group.

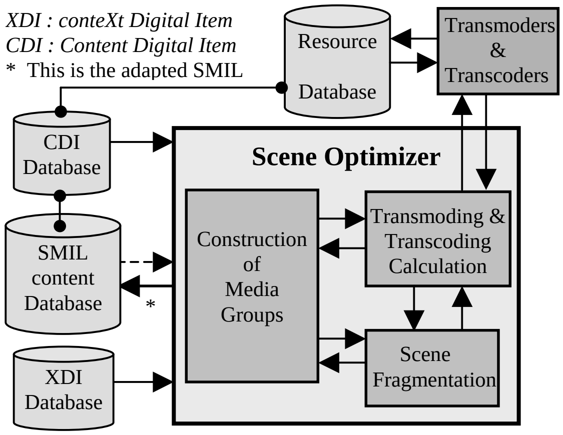
XDI : conteXt Digital Item
CDI : Content Digital Item
* This is the adapted SMIL
Resource Database
Transmoders & Transcoders
CDI Database
Scene Optimizer
SMIL content Database
Construction of Media Groups
Transmoding & Transcoding Calculation
*
Scene Fragmentation
XDI Database

Figure1. Scene-Adaptation Module Architecture

We perform this for all object groups. If possible, we can also integrate some consequent groups in one fragment. At the end, if no *impossible adaptation* happens, we end up by having a fragmentation into several scenes, which, in the adapted output SMIL scene, will be sequenced by, for example, a "click to see more" button in each scene fragment.

5.6. Resource adaptation

After the transcoding and transmoding calculations for each resource are finished, the resources requiring adaptation are transcoded or transmoded and then saved. The adapted SMIL refers to these saved resources. We use a set of transmoding and transcoding tools, which include some resource format conversions, visual media (image/video/text) resizing, video-to-image (and slideshow), graphics-to-video, and image-to-text transmodings.

6. Conclusions and perspectives

In this paper, we have proposed a methodology for semantic adaptation of synchronized multimedia presentations. By developing a semantic multimedia scene optimizer, we demonstrated that the expression of semantic information of a multimedia presentation is necessary to perform meaningful scene adaptation. Semantic adaptation of structured multimedia documents is a complex issue and needs to be addressed more completely. The order of complexity grows more significant as complex temporal dependencies are introduced between media objects of a scene. The focus of our future work is on enhancing our set of resource adaptors and on taking into account bandwidth limitations and the usage of MPEG-21 AQoS.

6. References

[1] I. Burnett et al. "Mpeg-21 goals and achievements", *IEEE MultiMedia*, October-December 2003, pp. 60-70.

[2] Rob Koenen, "Overview of the MPEG-4 Standard", http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm, March 2002.

[3] W3C, *Synchronized Multimedia Integration Language (SMIL) 1.0 Specification*, W3C Recommendation.

[4] R. Mohan, J.R Smith. Et al., "Adapting Multimedia Internet Content for Universal Access", *IEEE Transactions Multimedia*, March 1999, pp. 104-114.

[5] F. Rousseau et al., "User Adaptable Multimedia Presentations for the WWW", *Computer Networks*, 1999, pp. 1273-1290.

[6] J. Euzenat et al., "A Semantic Framework For Multimedia Document Adaptation", *IJCAI*, CA, US, 2003.

[7] Masahiro Hori et al., "Annotation-based Web Content Transcoding", *Computer Networks*, June 2000, pp. 197-211.

[8] Katashi Nagao et al., "Semantic Annotation and Transcoding: Making Web Content More Accessible". *IEEE MultiMedia*, 2001, pp. 69-81.

[9] M. Kimiaei A. and J-C. Dufourd, "Multimedia Adaptation by Transmoding in MPEG-21", *WIAMIS 2005*, Lisbon, Portugal, April 2004.