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 propose semantic annotation of multimedia scenes, expressing semantic information on each media object, as well as on the dependencies between all the media objects of the scene. We use these annotations in order to perform a semantic adaptation on multimedia presentations. We use our proposed description tools under the framework of MPEG-21, and we show that, in order to preserve the consistency and meaningfulness of the adapted multimedia scene, the adaptation process needs the semantic information of the presentation.

1. Introduction

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. Delivery of such diverse multimedia content to different contexts 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, user preferences, etc.), in order to be able to provide the end user with the optimal 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 standards that support and define frameworks for a multimedia content adaptation system. These standards, however, do not provide complete support for semantic adaptation. MPEG-21 [1] pays special attention to the support of resource (single media) adaptation to constrained contexts. It does not consider requirements for semantic adaptation of composed multimedia presentations. In this paper, we present our work on a semantic and context-aware multimedia adaptation framework based on MPEG-21. The adaptation of each resource and the whole multimedia presentation, takes into account the context constraints, and the semantic metadata of the multimedia content. Our semantic adaptation of multimedia presentations respects semantic relationships between the media objects.

In the remainder of this paper, we first give a short introduction to MPEG-21. We then provide a brief summary on the adopted approaches in the area of semantic adaptation of multimedia *scenes*. Next, will be described our methodology and an experimental implementation based upon it. Finally, the last section offers our conclusions and 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) and 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 medias. It also specifies how they should

0-7803-9332-5/05/$20.00 ©2005 IEEE

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. SMIL (*Synchronized Multimedia Integration Language*) [3], a W3C recommendation, is a high-level 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 shall 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 due to a bandwidth limitation, or lack of image support by the terminal, the adaptation engine should also remove the text caption of the image. This is not feasible without having semantic information of the scene.

Another simple example is a multimedia document with two images and two texts, each text giving explanation on one of the images. The display size of the user device is too small for the whole scene, even after maximum downscaling of the images. Fragmentation of the scene then becomes necessary. In this case, in order to keep the related image and text together in the same scene fragment, and to temporally sort the fragments in the correct order, the adaptation engine needs some semantic information of the scene.

As seen in these examples, a complete multimedia content adaptation 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 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 images, which is obtained from the original image [4]. F. Rousseau et al., propose solutions for the adaptation of multimedia presentations, however the solution remains incomplete, as it does not consider complete semantic dependencies between media objects [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 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 et al., propose semantic annotation and adaptation of HTML documents [8].

5. Multimedia Scene Semantic Adaptation (MSSA) framework

Our methodology for realization of a MSSA engine, is built upon five principle elements:

5.1. Content description

Explicit *physical* description of the content is very important in a multimedia adaptation framework. Although some of the characteristics of the content can be directly extracted from the content itself, such as its modality or format, there are some characteristics that shall be given explicitly, such as encoding parameters, semantic information (e.g. semantic key frames of a video, as opposed to “encoding” key frames), or the maximum spatial downscaling of a visual media, with which it is still logically visible. We call this latter *maxRRF:* maximum Resolution Reduction Factor.

We use MPEG-7 description tools for content description. For each media object, descriptors are wrapped up in a DID *Statement* element and attached to the *Resource* element referencing the corresponding media. We call CDI (Content Digital Item) the DID document containing 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. We use MPEG-21 DIA for the description of context. We call XDI (conteXt Digital Item) the DID document describing the usage context.

5.3. SID: Semantic Information Declaration

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 involving entity. We have defined XML schemes for the expression of semantic information of a multimedia scene. Our proposed SID (Semantic Information Declaration) descriptors are used by the adaptation engine to decide on the optimal type, nature and parameters values of the adaptation(s) that are to be applied to the scene. The information included in SID descriptors is categorized into three main parts: a) independent semantic information of each media object, b) semantic dependencies between media objects of the scene, and c) semantic preferences on scene fragmentation.

The first category describes, for each media object, its independent semantic information in the context of the scene, such as importance, role and *maxRRF*. For instance a key role media should not be removed or degraded under any circumstance. The second category includes: i) spatial dependencies; i.e. which media objects should be kept close together, ii) absolute semantic dependencies; i.e. which media object is or could be precondition or redundant to another media object, and iii) temporal dependencies; i.e. synchronization information between media objects. The third category describes preferences and priorities on the spatial and temporal fragmentation.

SID descriptors are given in the corresponding CDI, within a *Statement* element attached to a *Resource* element referencing the concerned multimedia scene.

5.4. Scene description

In our approach, we use SMIL 2.0 for describing scenes. However, our methodology is independent from this choice and can be applied to other multimedia description languages. The rationale behind this choice is that SMIL is a high-level scene description language. This makes it easier to perform manipulations on a SMIL scene than, say, an XMT scene. We map the media objects of the SMIL scene to objects in the DID instance.

5.5. Scene optimization

In this section we describe our proof-of-concept implementation of a *scene optimizer*. Figure 1 shows the architecture of our MSSA engine. The inputs are: description of context contained in an XDI, semantic information (SID) and *physical* description of content contained in a CDI, and the SMIL scene – which could alternatively be referenced through the CDI.

We consider the following rules: transmoding [9] is a pure modality conversion 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 considered to be smaller than the original scene layout size.

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 resources and the scene structure. Temporal synchronizations are also complicating factors.

The scene optimizer first verifies the modality and format support of the target device. Using the content description, it then attempts to convert (or replace) resources of the non-supported formats and modalities to (or by) other medias in other formats and/or modalities. If the attempt proves unsuccessful, the corresponding resources are simply removed 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* given in the CDI), the target display is still too small for 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 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 device display, we produce a scene fragment containing objects of this group. If not, using the *maxRRF* of each media object of this group, we calculate the minimum possible overall spatial size of this group. If this latter is smaller than the device display, the optimal downscaling for each resource is calculated based on its original size and its importance, so that the overall group resolution

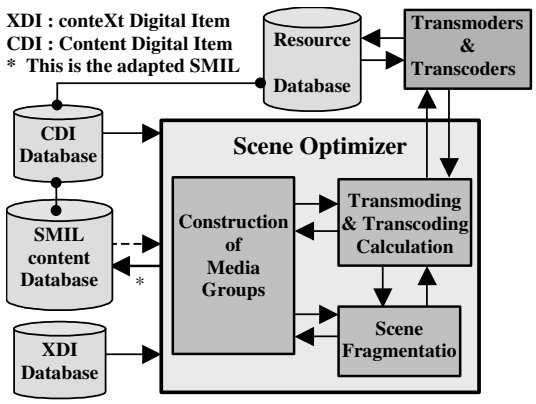

Figure1. Scene optimizer in MSSA architecture

becomes equal to target display size. If the minimum required space of the group – calculated based on *maxRRF*s of resources, is bigger than the display, we drop off redundant or low-importance medias, or replace them by a hyperlink, and then recursively redo the calculation for this new group until it is adapted to the available display size. This is done for all groups and if possible, consequent groups are integrated in one fragment. At the end, if no *impossible adaptation* happens, we will have several scene fragments, which will be sequenced by a "click to see more" button in each scene fragment.

5.6. Resource adaptation

After the transcoding/transmoding calculations are finished, the resources requiring adaptation are transcoded/transmoded. The corresponding media elements of the adapted SMIL will refer to these adapted resources. We used a set of transmoding/transcoding tools that include some media format conversions, visual media resizing/cropping, video-to-image, video-to-slideshow, graphics-to-video, graphics-to-image and image-to-text transmodings.

The most valuable advantage of MSSA engine is that it reduces the designing and authoring work. Instead of designing several layouts for different contexts, the author provides only the full version of his content and indicates its semantic information by means of simple XML descriptors. The MSSA engine then takes care of the rest by providing a personalized version of the original document for each request.

From an architecture point of view, if such engine is integrated into a client-server architecture, based on the capacities of the server, important response delays may be expected in case of numerous simultaneous adaptation requests. Distributed architectures are more appropriate in such cases. We also experienced that media position and resolution optimization calculations prove to be very complex for scenes with high number of media objects. Also for complex SMIL documents with numerous hyperlinks and timing dependencies, the programming task may become very heavy, nevertheless, the methodology remains valid. We also learned that multimedia documents, whose rendered layouts are pre-defined in an absolute manner, are not best adaptable, compared to device-independent multimedia documents. RIML [10] is an example of this kind of description languages, but only addresses the device-independent media positioning issues and does not take into account the between-media and media-independent semantic information.

6. Conclusions and perspectives

In this paper, we proposed a methodology for semantic adaptation of synchronized multimedia presentations. By developing a semantic multimedia scene optimizer, we showed that providing semantic information of a multimedia presentation is necessary to perform a meaningful scene adaptation. Semantic adaptation of structured multimedia documents is a complex issue. The order of complexity grows more significant as the number of media objects of the scene grows higher and as complex temporal dependencies are introduced between them.

Manipulation of multimedia documents that specify absolute spatial positions and dimensions for media objects, is quite complex. Hence, research on the subject of device-independent multimedia description languages, which address semantic adaptation of multimedia content, is an appropriate direction for future work in this area.

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", available on http://www.chiariglione.org, 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 Asadi and Jean-Claude Dufourd, "Multimedia Adaptation by Transmoding in MPEG-21", *WIAMIS 2004*, Lisbon, Portugal, April 2004.

[10] Consensus Project, "RIML Specification Document", available on http://www.consensus-online.org/publicdocs.