KNOWLEDGE-BASED AND SEMANTIC ADAPTATION OF MULTIMEDIA CONTENT

M. KIMIAEI ASADI AND J-C. DUFOURD

ComElec Department, ENST
46 rue Barrault,
75013 Paris, FRANCE
E-mail: mariam.kimiaei@enst.fr
E-mail: jean-claude.dufourd@enst.fr

In this paper, we present our work on two aspects of multimedia adaptation: knowledge-based single media adaptation and semantic adaptation of multimedia documents. For the former, we propose support of adaptation by direct hinting in the scope of MPEG-21. For the latter, which is the main part of the presented work, we introduce the description of semantic dependencies between media objects of a multimedia scene. The proposed description tools are also based on the framework of MPEG-21. In this paper, we aim to show that, in order to preserve the consistency and meaningfulness of the adapted multimedia scene, the adaptation peer 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 a large diversity of multimedia content to different types of user devices and environments is one of the major challenges of a multimedia delivery chain. The content creators, shall, therefore, take into account this point at the authoring level by creating adaptable content (i.e., by providing the necessary metadata for adaptation). The content delivery chains will also need to have enough information on the context of the usage environment (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 multimedia adaptation infrastructure is then needed to satisfy these requirements. Such an infrastructure will propose methods to express context constraints, as well as, content-related 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 adaptation system. In this paper, we will present our work on multimedia adaptation which is based on MPEG-21 [1]. The paper is divided into two main parts: single media (resource) adaptation and semantic adaptation of multimedia scenes. In single media adaptation, the process of adaptation is based on context constraints as well as metadata and direct hints provided by the author of the content. In this kind of adaptation, the media is considered solely (i.e. as mono media without any multimedia structured presentation) or independently of the multimedia composition (scene) in which it exists. Therefore, this kind of adaptation is not a complete solution. In the second part, which is the main part of the presented work, semantic adaptation of multimedia scenes addresses the adaptation of multimedia structured documents based on temporal, spatial and semantic relationships between the media objects.

The presented work is mostly done in the framework of two European IST projects: ISIS [2, 3] and DANAE [4].

2. Single Media Adaptation by Direct Hinting

MPEG-21 has a set of detailed and complete adaptation and descriptions tools for single media adaptation based on end user constraints and preferences. Adaptation based on direct recommendations, suggestions or hints of the author of the content are not yet fully supported by MPEG-21. The creator of a multimedia content, based on his knowledge on the resource, may have specific hints or more generally, metadata, for specific types of adaptation of his resource. We have proposed extensions to MPEG-21 DIA (Digital Item Adaptation) [5, 6] to support adaptation based on author direct hints. These contributions deal mostly with the support of adaptation by on-line *transmoding* based on the information provided by the author of the resource.

Transmoding is defined to be the adaptation of a digital media by (on-line) transformation of its original *modality* to another modality [7]. We consider the following five principal modalities, along with several sub-modalities (as shown parenthetically): Video, Audio (Audio2D, Audio3D, Speech), Image, Graphics (Graphics2D, Graphics3D) and Text.

The need for support of adaptation by transmoding was investigated and then validated in the use-cases of the European ISIS project. In ISIS we encountered real-use-cases of on-line transmoding for which, in order to perform the resource adaptation, we needed to have access to some transmoding parameters. We, therefore, defined a description tool for transmoding, as an extension to MPEG-21 DIA, to facilitate the production of metadata that describes resource-adaptation by transmoding [7]. The transmoding hints include the descriptions of the most general transformation parameters, i.e., the descriptors are based on no particular underlying algorithm. Some of these parameters are, for example, the key frames of a video, and their relative importance, for a video to image (or slide-show) transmoding.

We used the transmoding description tool, to integrate adaptation cases of (on-line) transmoding type by direct hinting into ISIS [7].

Direct hints and recommendations of the author of the content could be quite helpful to guide or sometimes enable the adaptation process. As explained in the next section, we use direct hints for adaptation of single media resources in the context of a multimedia scene.

As the continuation of this work on description tools for direct hinting, and under the framework of DANAE project, we are also working on the support of expression of author direct hints for adaptations of transcoding type. For example, for a visual media resizing, one of the parameters could be the maximum resolution reduction factor that the author recommends. As you will see in the next section, we use this parameter (as a direct hint in a transcoding descriptor attached to a visual media resource in the DID instance [8]) in order to calculate the limits of spatial downscaling of this visual resource.

3. Semantic Adaptation of Multimedia Scenes

A multimedia scene is a synchronized multimedia presentation that integrates multiple static, or continuous medias. 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. The MPEG group has developed XMT and BIFS (*BInary Format for Scenes*) which are description languages for MPEG-4 scenes [9]. SMIL (*Synchronized Multimedia Integration Language*) [10], a W3C recommendation, is a specification language with temporal functionalities.

When adapting a multimedia presentation, in order to preserve the consistency and meaningfulness of the adapted scene, the adaptation peer 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 non-supporting of image modality by terminal, the adaptation engine should also remove the caption of the image. This is not feasible without having the semantic information of the scene which includes the semantic relationship between the two media objects (image and text caption).

Another simple example can be illustrated by a multimedia document with two images and two texts, each giving explanation on only one of the two images. Let's assume that the display size of the user device is too small to display the whole scene, even after maximum downscaling of the images. A fragmentation of the scene may then be 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.

Z. Lei et al. discuss different issues of a general context-based adaptation framework in [11]. Mohan and Smith proposed a framework for the adaptation of multimedia documents [12], in which, the single media adaptation is done by selecting the optimum version of a media among pre-transcoded and pre-transmoded versions of it. The presentation adaptation is based on some semantic information (mainly on the purpose of image media objects) that is obtained from the original image object [13]. F. Rousseau et al., also propose solutions for the adaptation of multimedia presentations that remain incomplete from the semantic point of view [14]. J. Euzenat et al., present solutions for adaptation of multimedia documents along their temporal dimension [15].

3.1. *Semantic Information Description*

We have defined XML [16] schemes as extensions to MPEG-21 DIA for the expression of semantic information of a multimedia scene. Like conversion descriptors, these descriptors are attached to a multimedia resource in the DID instance. The SID (Semantic Information Description) descriptors contain information provided by the author of the multimedia scene and are then used by the decision making engine to decide on the type and nature of the adaptation(s) to be applied to the scene. The information included in SID descriptors is categorized into two main parts: semantic dependencies between media objects of the scene, and semantic preferences on scene fragmentations. The former includes spatial dependencies (which media objects should be kept close together), absolute semantic dependencies (which media object is, or could be, a precondition, or a redundant for another media object) and temporal dependencies (synchronization information between media objects), while the latter describes preferences (priorities) on spatial and temporal fragmentation. SID also describes, for each media object, its independent semantic information in the context of the scene, such as its importance, role (if it has a basic role, i.e. should not be, in any case, removed or degraded), etc.

3.2. *Scene Description*

In our approach we use SMIL for describing scenes, however, this could be done with any other multimedia description language. The reason behind this choice, is that SMIL is a high level scene description language, therefore, manipulating (performing adaptations on) a SMIL scene, compared to, for example, a XMT scene, is easier. We also map the media objects, which are present in the SMIL scene, to media objects in the DID instance.

For our purpose, and in order to reduce the complexity of the work, we have considered templates for the SMIL scenes that we use.

3.3. *Scene Optimiser Algorithm*

In this section we describe our *scene optimiser* algorithm. To perform the *optimal* decision-making, we define a set of rules and assumptions:

- Scene fragmentation is preferred to resizing of single media objects. This means that *scene adaptation* is preferable to *media adaptation*, in case the overall adaptation is possible only by scene fragmentation and without any single media adaptation.
- Single media transmoding is a pure modality conversion and has no effect on the spatial size (resolution) of a visual media.
- Only low-importance and redundant medias can be removed from the scene.
- Maximum spatial downscaling of each visual media is given. This will be called *maxRRF* (maximum Resolution Reduction Factor) and could be given for example in a transcoding descriptor associated to the related media resource in the DID instance (CDI: Content Digital Item).
- The display size of the target device is always smaller than the display size for which the scene has been originally authored.

Having defined the above policies, the algorithm of our *scene optimiser* module is still quite complicated. The reason for this complication is that we have to, simultaneously, optimise both the single media and the scene adaptations. Temporal synchronizations between media objects are also complicating factors. Here, in order to make this easier to understand, we simplify our *scene optimiser* algorithm as follows:

1. Verification of the modality support of the target device and then removal of the media objects of the non-supported modalities.
2. If possible, replacing these medias by other medias in other modalities using the transmoding descriptor which is associated to media resources in CDI. And if not possible, removing them from the scene. We have to keep in mind that in every step of the optimising, when a *basic role* media is to be removed, adaptation is considered to be impossible and the optimising process is cut, we call this, an *adaptation impossible* case.
3. Checking the target display size. If it is smaller than the layout of the scene, then, based on the information given in SID descriptors, we construct groups of

media objects so that all media objects which are semantically related to each other, stay in the same group. We then sort these object groups by their timing priorities (an information which is given in the SID descriptor of each object).

4. Starting with the object group of the highest timing priority, we calculate the overall spatial size (resolution) of the group for each group. In case this is less than the size of the target device, we produce a scene fragmentation containing objects of this group. And if not, using the *maxRRF* of each media object of this group, we calculate the minimum possible overall resolution of this group. Then:
   - If this is smaller that the target device, 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, from the group, and we then redo the step 4 for this new group.

We perform step 4 for all object groups. If possible, we can also integrate some consequent groups in one fragmentation. At the end if no *adaptation impossible* happens, we end up by having several scene fragmentation, which, in the adapted output SMIL scene, will be sequenced by, for example, a "click to see more" button in each scene fragmentation.

3.4. *Media Adaptation*

After the optimising, i.e. the transcoding and transmoding calculations, is finished, based on the performed calculations, the object medias (those who need to be adapted) will be transcoded or transmoded and then saved. Evidently, the adapted media objects in the final adapted SMIL scene refer to these saved resources.

Figure 2 shows the architecture of our scene adaptor module. XDI: conteXt Digital Item, is a DID & DIA instance (in DIDL language) containing the information on the context of the usage: terminal display size, terminal supported modalities and etc. CDI: Content Digital Item is also a DID & DIA instance (in DIDL language) which contains the SID information for all media objects of the scene and the other content-related information, such as transcoding and transmoding descriptors.

We use a set of transmoding and transcoding tools, which include visual media (image/video/text) resizings, video to image (and slideshow), graphics to video, and image to text transmodings.

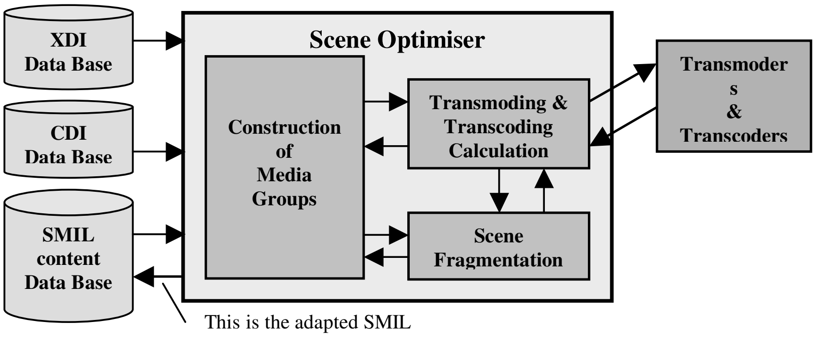

Figure2. The architecture of the scene-adapting module.

4. Conclusions and Perspectives

Our work on knowledge-based multimedia content adaptation distinguishes between two different kinds of multimedia content adaptations: single media adaptation and multimedia presentation adaptation. We have proposed solutions for single media conversion by direct hinting. We have also proposed a generic solution for semantic adaptation of synchronized multimedia presentations. We have specified means of expressing semantic information of a multimedia structured document in generic multimedia presentations. We showed that the expression of semantic information on media objects of a multimedia presentation, is necessary for performing a meaningful scene adaptation. Adaptation of structured multimedia documents, based on semantic information, is quite a difficult question that needs to be addressed more completely. The complication is yet more significant when we introduce complex temporal dependencies between objects of a scene. Our perspectives on this work, are in a first step, to further work on the simultaneous optimisation and adaptation of media objects and scene itself, and in a second step to consider bandwidth limitations and the usage of MPEG-21 AQoS in our single media adaptation.

References

1. I.Burnett, R. Van de Walle, K. Hill, F. Pereira « Mpeg-21 goals and achievements », *IEEE MultiMedia*, vol. 10,No. 4, p. 60-70 (October-December 2003).
2. http://isis.rd.francetelecom.com/

3. K. Kamyab, et al. « ISIS: Intelligent Scalability for Interoperable Services », Conference on Visual Media Production CVMP 2004, London (March 2004).
4. http://danae.rd.francetelecom.com/
5. MPEG MDS Group, Multimedia framework (MPEG-21) – Part7: Digital Item Adaptation, (Final Committee Draft), ISO/IEC JTC 1/SC 29/WG 11/N5845 (July 2003), http://www.chiariglione.org/mpeg/working_ documents.htm
6. A. Vetro, « MPEG-21 Digital Item Adaptation: Enabling Universal Multimedia Access », *IEEE Multimedia*, vol. 11, No. 1, p. 84-87 (January-March 2004).
7. M. Kimiaei Asadi and J-C. Dufourd, Multimedia Adaptation by Transmoding in MPEG-21, 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Lisbon, Portugal (April 2004).
8. MPEG Group, ISO/IEC 21000-2:2003 Information technology - Multimedia framework (MPEG-21) - Part 2: Digital Item Declaration (2003).
9. Rob Koenen, Overview of the MPEG-4 Standard ISO/IEC JTC1/SC29/WG11 N4668 (March 2002), http://www.chiariglione.org/ mpeg/standards/mpeg-4/mpeg-4.htm
10. W3C, Synchronized Multimedia Integration Language (SMIL) 1.0 Specification, W3C Recommendation, http://www.w3.org/TR/1998/REC-smil-19980615.
11. Z. Lei, N. D. Georganas, « Context-based Media Adaptation in Pervasive Computing », Proceeding of Canadian Conference on Electronic and Computer Engineering (CCECE'2001), Toronto (May 2001).
12. R. Mohan, J.R Smith. and C.S. Li, « Adapting Multimedia Internet Content for Universal Access », *IEEE Transactions Multimedia*, vol. 1, No. 1, p. 104-114 (March 1999).
13. S. Peak and J. R. Smith « Detecting Image Purpose in World-Wide Web Documents », Symposium on Electronic Imaging: Science and Technology - Document Recognition, IS&T/SPIE (1998).
14. F. Rousseau, J.A. García-Macías, J. Valdeni de Lima, A. Duda, « User Adaptable Multimedia Presentations for the WWW », *Computer Networks*. vol. 31, No. 11-16, p. 1273-1290 (1999).
15. J. Euzenat, N. Layaïda, V. Dias, « A semantic framework for multimedia document adaptation », Proceeding of 18th International Joint Conference on Artificial Intelligence (IJCAI), Acapulco (MX), San-Mateo (CA US), p. 31-36 (2003).
16. F. Yergeau, J. Cowan, T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, Extensible Markup Language (XML) 1.1, W3C Recommendation (February 2004).