A Semantic-based Multi-modal Utility Approach For Multimedia Adaptation

Martin Prangl[1]**, Hermann Hellwagner**[1]**, Tibor Szkaliczki**[2]

[1] Klagenfurt University, Department of Information Technology
9020 Klagenfurt, Austria
[2] Computer and Automation Research Institute of the Hungarian Academy of Sciences
eLearning Department, 1111 Budapest, Hungary
Dennis Gábor College, 1037 Budapest, Hungary

Abstract Content adaptation is an important issue of multimedia frameworks in order to achieve Universal Multimedia Access (UMA), i.e., to enable consumption of multimedia content independently of the given resource limitations, terminal capabilities, and user preferences. The Digital Item Adaptation (DIA) standard, one of the core specifications of the MPEG-21 framework, supports content adaptation considering a wide range of networks, devices, and user preferences. Most adaptive multimedia frameworks targeting the UMA vision do not consider utility aspects in their adaptation decisions. This paper focuses on a generic multi-modal utility model for DIA currently under design and evaluation, that aims to enhance the multimedia experience for the client. Our proposed model is able to take the semantics and the perceptual features of the content as well as the clients' specific utility aspects into account. Based on a detailed analysis of these constraints, we will show how the model reacts on individual input data. Finally, we will discuss results of the multi-modal decision taking process according to a few use case scenarios.

1 Introduction

Multimedia services over computer networks are becoming widespread. The multimedia content can be delivered to different terminals such as desktop PCs, PDAs, and mobile phones. There has been a significant amount of research recently on the adaptation of multimedia contents to the actual usage context to ensure Universal Multimedia Access (UMA). In many situation, the clients are typically unable to receive large audio-visual data volumes in original quality because of resource limitations. The following question has become crucial: "How to adapt multimedia data in order to provide the best user perceived utility?"

To answer this question, physical issues such as terminal capabilities, network characteristics etc. have to be considered. However, the quality of the adaptation significantly depends on the information of the content as well. For example, it would be better to adapt an action video in the spatial domain than in the temporal domain. As a consequence, the user would get a smaller video window but he/she would still be able to fully enjoy rapid motion in action scenes. Therefore, the semantics of the content should be taken into consideration in the adaptation decision process. Moreover, especially in utility based adaptation frameworks, the semantic experience of a content should be optimized under given resource limitations. In this paper, we will introduce a multi-modal adaptation decision model for DIA [1], which uses detailed perceptual quality information and semantic quality estimation.

When considering quality in the multimedia area, we have to distinguish between its perceptual part and its semantic part [2]. The perceptual quality (PQ) is a metric about how a user perceives the content, and refers to the human visual system (HVS). For example, a smooth video has a higher perceptual quality than a flickering one. The semantic quality (SQ) on the other hand includes the designated information that the medium should convey to the user, e.g., the semantic content of a news report or the motion aspect of an action video [3]. Furthermore, there is a big difference between quality and utility in the area of multimedia applications. The term quality is mostly used to refer to the perceptual quality whereas utility is a metric of satisfaction of the end user consuming this content. In this paper, the term utility consists of both the perceptual and the semantic part of quality for the given content.

So called *cross-modal utility models* are used to estimate the total utility of a media stream consisting of two or more modalities, e.g., video and audio. The total utility can be interpreted as a function which depends on the uni-modal utilities of the elementary streams themselves. In case of two modalities, namely video and audio, the total utility U can be defined as $U = f(U_V, U_A)$. U_V represents the video utility and U_A the utility of the audio stream. In the literature, there are some implementations of such a function; see, e.g., [4] for a discussion. All these implementations rely on adding the weighted uni-modal perceptual qualities, a multiplicative term (multiplication of uni-modal qualities), and specific constants in order to fit the subjective impressions of a group of test persons. The result of a detailed analysis of this approach [4] is that the implementation of the model itself as well as the weights and constants are strongly dependent on the genre and the subjects participating in the test. For this reason, we see the lack of a more generic model for estimating the total multi-modal utility which can be used for any genre and which takes into account the individual client's preferences.

In our opinion, an approach for defining such a generic model has to start from the other direction. We avoid subjective perceptual testing because it is expensive and time-consuming. Rather than giving a group of users a set of content variations for subjective testing, the individual user should be asked for his/her personal utility aspects. From these, such a generic model should be configured by fitting the model parameters to satisfy his/her individual preferences and utility concept. In case of this utility model, high total utility should indicate high subjective perceptual quality as well.

The next section gives an overview of the proposed multimedia framework for cross-modal utility modelling. Then, the utility model is introduced in detail. Based on use cases, we will show how it is possible to map high level user preferences and usage environment parameters to the semantic quality of the media stream.

2 Multimedia Framework with Cross-modal Utility Modelling

Figure 1 shows the concept of the proposed approach and its integration into a multimedia framework. The given user preferences and the genre (influencing SQ) have to be known for configuring our generic utility model which is used by the adaptation decision taking engine (ADTE) [5]. This input information is mapped to specific model parameters which we call *high level adaptation parameters*, discussed in Section 3. The individually configured model additionally needs to know the PQ of all deliverable content variations. Based on this information, the total utility U of all deliverable A/V variations can be estimated. Having available the utility of each deliverable A/V variation, the information about the required resources (e.g., bit rate), and the information about the resource limitations on the client and server sides (e.g., the available bandwidth, battery status, or CPU power), the ADTE is then able to estimate the optimal adaptation strategy for the individual content request [5]. This optimal adaptation decision is expressed by a set of parameters which we call *low level adaptation parameters*. They define an A/V media stream variation by its features (e.g., frame rate, spatial resolution, sample rate). Based on these target features, the adaptation engine (AE) finally performs the adaptation step on the original content. Finally, the produced variation, fitting the individual user preferences and environment and providing the best possible utility under given conditions, can be delivered for consumption to the requesting client.

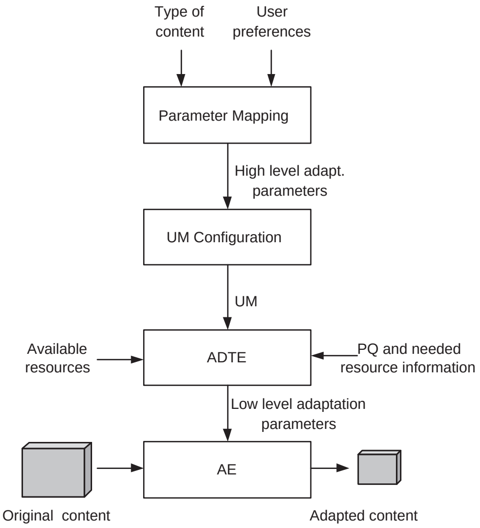

Fig. 1 Overview of integrated multimedia framework with cross-modal utility modelling.

3 Utility Model

The basis of the proposed model is that the total utility U_E of an elementary stream E can be split up into a perceptual part and a semantic part [2] as follows:

$$U_E = s \times PQ + (1 - s) \times SQ$$

where $s\epsilon[0..1]$ denotes a weight that indicates how much influence PQ has on the total utility. Because SQ is the most important part indicating how the user receives the designated content information, we define $s = 0.3$ in our model implementation. Note that PQ and SQ are normalized, i.e., in the range between 0 (worst) and 1 (best value). In the cross-modal case, we have to merge the utilities of the video and audio parts as follows:

$$\begin{aligned} U = \alpha \times [s \times PQ_A + (1 - s) \times SQ_A] + \\ (1 - \alpha) \times [s \times PQ_V + (1 - s) \times SQ_V]. \end{aligned} \quad (1)$$

PQ_A and PQ_V are representing the perceptual quality of the audio and the video part, respectively, and SQ_A and SQ_V represent the corresponding semantic qualities. A multiplicative term and an additive constant, as used in perceptual cross-modal quality modelling [4], is omitted in our multi-modal utility approach. The reason is that we do not use a regression analysis based on subjective tests. Furthermore, the multiplicative cross-modal perceptual influence would be negligible in our case because our model is strongly bound on additive uni-modal semantic aspects. α denotes the importance weight of the audio utility. It represents a high level adaptation parameter. All high level parameters are directly depending on the genre and individual user preferences. These user preferences as well as the usage environment can be easily described by MPEG-21 DIA Usage Environment Descriptors (UED) [1] for interoperable exchange. For example, in the case of a newscast, the importance of the audio part would be higher than the video part, resulting in a high value of α. A further use case would be that the user is hearing impaired or deaf. Then the value of α would be low or zero.

For perceptual video quality estimation, we use an objective measure, the peak signal to noise ratio (PSNR), which is widely used in image and video analysis. There are several perceptual video quality metrics, designed to better fit the HVS than PSNR. However, the result of a detailed comparison of these metrics, which was done by the Video Quality Experts Group (VQEG), shows that the metrics do not differ significantly in statistical results [6]. As an example, Figure 2 shows the normalized PSNR values for an MPEG-4 encoded high motion action video with fixed spatial resolution and varying frame rate and quantization parameter. For audio PQ estimation, we refer to [7] due to space limitations.

It is not suitable to do PQ estimation online due to the high computational requirements. However, the offline PQ results can be provided by the MPEG-21 DIA Adaptation QoS (AQoS) descriptor [1]. The resulting PQ values for frame rates below 18 fps are nearly equal (Figure 2). This fact implies the question: What is the "best" adapted variation for the end user? The answer is that this depends on the semantics of the content, the information which the user should receive by consuming the media stream. This semantics can be derived from the genre and the corresponding importance of its individual low level adaptation parameters. For example, in case of an action video delivered under bandwidth limitations, the semantic experience would be higher if the video were adapted in the spatial domain than in the temporal domain; i.e., the spatial resolution should be reduced and the frame rate of the original video should be kept intact. This adaptation step would result in a smaller window, but retain high motion in the video.

This consideration leads us to the following definition of the relative value of the semantic content of an individual elementary media stream (SQ_E):

$$SQ_E = f(W, F) \tag{2}$$

where W is a set of individual high level parameters (user and genre specific) and F represents a set containing data indicating the degradation of each feature in the stream (content variation specific).

The definition of the semantic video quality SQ_V is given in Eq. (3). The high level parameters w_{Fv}, w_{Sv}, and w_{Qv} act as importance weights of the video stream features. Note that the unique stream features of the video variation are nothing else than the low level video adaptation parameters. fr represents the frame rate, $height$, and $width$ the spatial resolution and q the quantization parameter of the video variation. q_{max} represents the codec specific maximum quantization value. fr_{orig}, $height_{orig}$, and $width_{orig}$ are constants representing the corresponding features of the original video stream. Figure 3 shows the graphical representation of SQ_V for one spatial resolution of the video stream and varying frame rate fr and quantization parameter q and for two different sets (W) of high level parameters. It explains the effect of applying different sets of high level parameters. The resulting SQ_V points form a plane in the space, where the high level parameters act as weights defining the slope of the plane. The original video stream ($q = 1, fr = 25$) has the highest SQ value whereas the worst value results by the most degraded variation ($q_{max} = 31, fr = 1$).

$$\begin{aligned} SQ_V = w_{Fv}\frac{fr}{fr_{orig}} + w_{Sv}\frac{height}{height_{orig}}\frac{width}{width_{orig}} + \\ +w_{Qv}(1 - \frac{q-1}{q_{max}}) \end{aligned} \tag{3}$$

$$\begin{aligned} w_{Fv}, w_{Sv}, w_{Qv}\epsilon[0..1], w_{Fv} + w_{Sv} + w_{Qv} = 1 \\ fr \leq fr_{orig} \\ height \leq height_{orig} \\ width \leq width_{orig} \\ q\epsilon[1..q_{max}]. \end{aligned}$$

The definition of the semantic audio quality again relies on a weighted approach of its modality features like sample rate (sf), encoding bit rate (abr) and the number of channels ($achan$) which is given in Eq. (4). The high level parameters w_{Sa}, w_{Ba}, and w_{Ca} act as importance weights of the audio stream features.

$$\begin{aligned} SQ_A = w_{Sa}\frac{sf}{sf_{orig}} + w_{Ba}\frac{abr}{abr_{orig}} + \\ +w_{Ca}\frac{achan}{achan_{orig}} \end{aligned} \tag{4}$$

$$\begin{aligned} w_{Sa}, w_{Ba}, w_{Ca}\epsilon[0..1], w_{Sa} + w_{Ba} + w_{Ca} = 1 \\ sf \leq sf_{orig} \\ abr \leq abr_{orig} \\ achan \leq achan_{orig} \end{aligned}$$

In order to show the reaction of the proposed multimodal utility model, we fitted appropriate high level parameters according to individual use case (UC) scenarios as shown in Table 1. On the right hand side of the table, the low level parameters with the highest total utility values are given. UC1 represents the delivery of an action

Table 1 Changing the preferences of different modalities.

UC	α	w_{Fv}	w_{Sv}	w_{Qv}	w_{Sa}	w_{Ba}	w_{Ca}	*bw limit*	res	fr	q	sr	abr	achan
1	0.5	0.6	0.2	0.2	0.3	0.3	0.4	200	128x96	25	10	48	160	2
2	0.3	0.1	0.5	0.4	0.4	0.4	0.2	200	704x576	5	16	48	64	2
3	0	0.6	0.2	0.2	dc	dc	dc	200	128x96	25	3	-	-	-

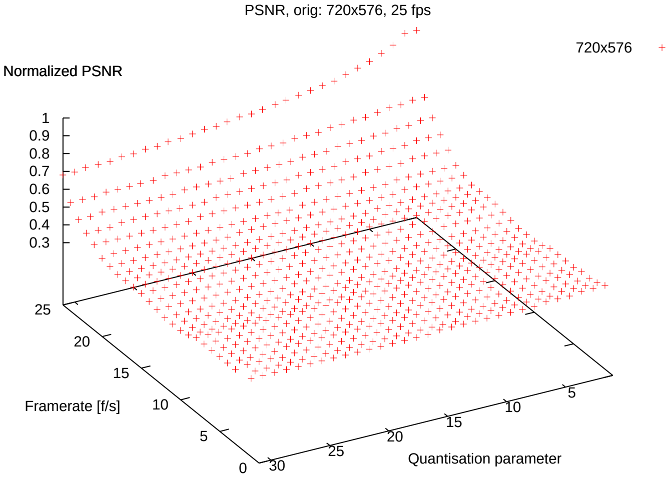

Fig. 2 PQ of all variations of high-motion MPEG-4 video for fixed spatial resolution, varying frame rate and quantization.

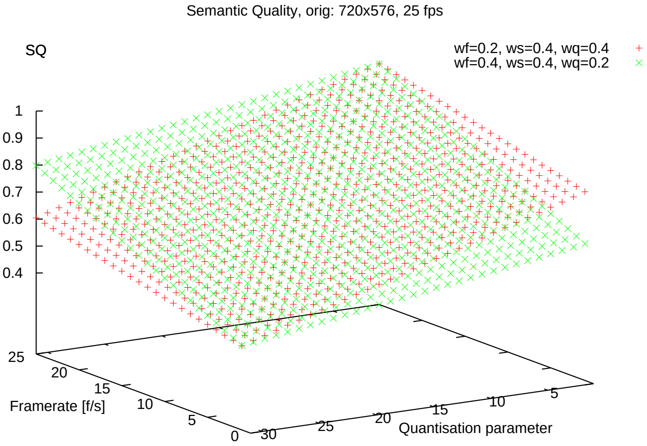

Fig. 3 SQ of video for varying frame rate and quantization.

video, UC2 a nature video, and UC3 sports (soccer) for hearing impaired users, to be delivered under bandwidth limitations. We selected relatively large weights for the frame rate and the spatial resolution in case of UC1 and UC2, respectively. The high level video parameters of UC3 are the same as in the case of UC1 but the audio parameters are irrelevant. All original A/V streams need a bit rate of 5–6 Mbps with the following features: $res = 720x576, fr = 25fps, q = 1, sr = 48kHz, abr = 160kbps, achan = 2$. We applied the adaptation decision taking algorithms presented in [4] to determine the low level parameters. In case of UC1, a variation with the highest frame rate resulted in the maximum utility. In UC2, the maximum utility was reached at the highest spatial resolution under the same bandwidth limitations. In case of UC3, the quantization of the video is improved as compared to UC1 at the cost of discarding the audio data.

4 Conclusions and Further Work

We presented a generic multi-modal utility model for multimedia content adaptation that is able to consider the usage environment as well as different genres. Applying this model to the adaptation decision taking process yields a better multimedia experience for the client.

Further experimental work will be performed to appropriately fit the high level parameters to different usage environments and genres. A recommendation system will be devised and implemented that predicts these parameters for various usage and content type scenarios, asks for and considers user satisfaction, and refines the parameters according to the users' judgements.

Acknowledgement: This work was in part supported by the Mobile Innovation Center, Hungary.

References

1. A. Vetro and C. Timmerer, "Digital Item Adaptation: Overview of Standardization and Research Activities", *IEEE Trans. on Multimedia*, vol. 7, no. 3, June 2005.
2. T.C. Thang, Y.J. Jung, and Y.M. Ro, "Modality Conversion for QoS Management in Universal Multimedia Access", *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 152, pp. 374-384, June 2005.
3. T.C. Thang, Y.J. Jung, and Y.M. Ro, "Semantic Quality for Content-Aware Video Adaptation", *Proc. IEEE MMSP'2005*, Shanghai, China, Oct. 2005.
4. D.S. Hands, "A Basic Multimedia Quality Model", *IEEE Trans. on Multimedia*, vol. 6, no 6., Dec. 2004.
5. M. Prangl, H. Hellwagner, and T. Szkaliczki, "Fast Adaptation Decision Taking for Cross-modal Multimedia Content Adaptation", *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Toronto, Canda, July 2006, accepted.
6. M.H. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality", *IEEE Trans. on Broadcasting*, vol. 50, no. 3, Sept. 2004.
7. B. Feiten, I. Wolf, E. Oh, J. Seo, and H.-K. Kim, "Audio Adaptation According to Usage Environment and Perceptual Quality Metrics", *IEEE Trans. on Multimedia*, vol. 7, no. 3, June 2005.