MODALITY CONVERSION FOR UNIVERSAL MULTIMEDIA SERVICES

Truong Cong Thang, Yong Ju Jung, Jae Wook Lee, Yong Man Ro

Multimedia Group, Information and Communication University (ICU), South Korea

ABSTRACT

Modality conversion (sometimes called transmoding) currently emerges as an important issue in Universal Multimedia Access. The decision on modality conversion is affected by various factors, such as terminal capability, user preferences, surrounding environment, etc. Here, we consider modality conversion under the constraint of available resource. Intuitively, when content scaling cannot provide the acceptable QoS, modality conversion may be the good choice to maintain the quality. From the QoS point of view, two important questions in modality conversion are "at which resource constraint should the current modality be converted?" and "what is the destination modality?" That is, knowing the conversion boundaries between modalities is crucial for a seamless modality conversion. In this paper, we present a systematic approach to help answer these questions.

1. INTRODUCTION

Universal Multimedia Access (UMA) is currently a new trend in multimedia communications [1]. In a UMA system, content adaptation is the most important process to cope with various constraints of terminal and network. Content adaptation has two major aspects: one is *modality conversion* that converts content from one modality to different modalities, the other is *content scaling* that changes the bitrates (or qualities) of the contents without converting their modalities. So far, most researches on content adaptation have dealt with content scaling.

The modality concept of multimedia content is actually quite broad. It can be considered from the human senses such as visual, auditory, tactile, etc. These modalities have been tackled for a long time in the field of human-computer interface. Modalities can be also derived from different modes of content coding (e.g video, image, graphics for visual sense). Even, different coding formats (e.g. GIF, JPEG for image) are sometimes referred to as the modalities or sub-modalities. MPEG-7 have defined various classification schemes to describe these "hierarchical" modalities (e.g. ContentCS, AudioCodingCS, GraphicsCodingCS, etc).

In the evolution of UMA, modality conversion currently appears to be an important demand. There are various conditions that may affect the decision on modality conversion. In our opinion, they can be grouped into four main factors. The first factor is the *modality capability*, which is the support for user's consumption of certain modalities. This factor can be determined from the characteristics of terminal (e.g. text-only pager), or surrounding environment (e.g. a too noisy place). The second factor is the *user preference* that shows user's levels of interest to different modalities. The third factor includes the *resource constraints*, for example the terminal can support video modality but at some point the connection bitrate is not enough to play the video content online. The fourth factor is the *semantics* of the content itself. For instance, between an interview video and a ballet video, the provider would be more willing to convert the former to a stream of text.

MPEG-21 Digital Item Adaptation (DIA) provides various Usage Environment description tools to help determine the modality capability, and the Conversion-Preference tool for users to personalize their use of content modalities [2].

Currently, modality conversion is often carried out only when some modality is not present in the modality capability. In this paper, modality conversion is considered mainly with the resource constraint factor. Intuitively, given some resource constraint of terminal/network, the provider will (down) scale the contents to meet the constraint while still providing the best possible quality to the user. However, in some cases, the quality of the scaled contents is unacceptable or not as good as that of a substitute of a different modality. A possible solution for this problem is to convert the contents into other modalities. For example, when the connection bitrate is too low, sending a sequence of "important" images would be more appropriate than streaming a scaled video of low quality. This is a typical case of conversion from video modality to image modality. From the QoS point-of-view, two most important questions for modality conversion are:

- "At which resource constraint should the current modality be converted?" and,
- "What is the destination modality?"

The goal of our work is to help the adaptation system answer these questions when the terminal can support the modalities but the resource constraint is limited. The paper is organized as follows. In section 2, we describe

the basic ideas of modeling modality conversion. In section 3, modality conversion for networked video service is explored, in which we consider three possible destination modalities, namely image, audio, and text. Finally the conclusions are discussed in section 4.

2. MODELING CONTENT SCALING AND MODALITY CONVERSION

The process of content scaling can be represented by some "rate-quality" curve, which shows the quality of the scaled content according to the bitrate (or any resource in general). The recent trend in UMA is to use this rate-quality curve to enable the automation of content scaling [3][4]. MPEG-21 DIA provides several description tools (AdaptationQoS) for this kind of modeling [2].

We see that different modalities have different characteristics and their qualities may be measured in different dimensions (e.g. PSNR, MOS), so the rate-quality curve should be computed *within a particular modality*. Still, we need to compare the qualities of different modalities, so as to find the conversion boundary. The Overlapped Content Value (OCV) model, first mentioned in [5], helps clarify the relationship among content value (quality), resource, and modalities.

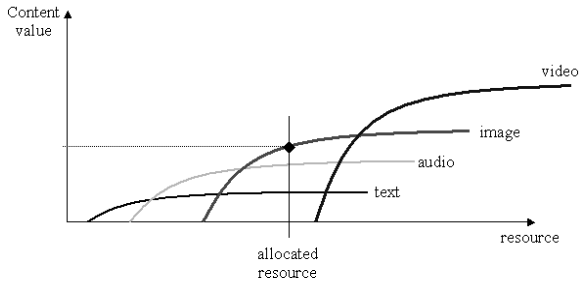


Fig. 1: Overlapped content value model of a content.

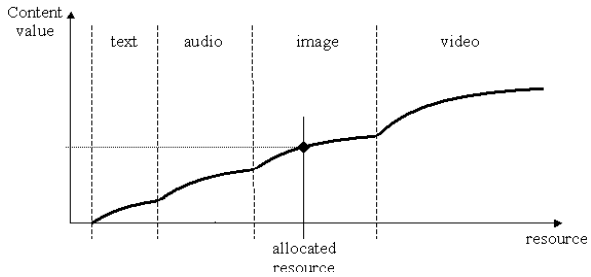


Fig. 2: The final content value function of the content.

Fig. 1 shows the example OCV model of a video content, which consists of the rate-quality curves of different modalities (called *modality curves*). These modality curves are normally non-decreasing and can be provided manually or automatically. We can see that the intersection point of the modality curves represents the conversion boundaries between modalities. Denote $VM_j(R)$ the rate-quality curve of modality j of the content, j=1...J, where J is the number of modalities of the content and j=1 is the index of the original modality; R is the resource. $VM_j(R) \geq 0$ for all j=1...J. The content value function can be represented as follows:

$$V = max\{w_j \cdot VM_j(R) \mid j=1...J\} \quad (1)$$

where w_j is the scale factor of modality j of the content.

Fig. 2 shows the final content value function and the conversion boundaries of the content. Based on this model, we can quantitatively make the decision on modality conversion as well as content scaling, so as to maintain an acceptable QoS.

As suggested by (1), to harmonize the different modalities within a model, the provider assigns an appropriate *scale factor* w_j to each modality, so as the content values of different modalities reflect their relative importance and have a common unit. Note that the OCV model is applicable to both offline and online adaptation. This model can also be used as the underlying basis to support user preference on modalities [5].

By the proper estimation of content value and scale factors for modalities, we can put different modality curves into an OCV model, and then determine the conversion boundaries between the modalities. However, such estimation is not easy; it depends on the modalities and semantics of the contents. For example, let's consider the conversion from video to text. In case of concert video, the scale factor for text should be very small; while in case of interview video, the scale factor for text may be much higher. Anyway, if the provider carries out careful subjective tests for the content, he can reasonably select an appropriate scale factor for each modality.

3. A CASE STUDY OF MODALITY CONVERSION IN NETWORKED VIDEO SERVICE

In this section, we will explore the possibility of modality conversion for a video streamed through a network to the terminal. We employ a Foreman MPEG-4 video stream, having the bitrate of 119.3Kbps, frame rate of 25fps and length of 300 frames. The operations of content scaling and modality conversion in our experiment are carried out offline. A number of content versions of video, image, audio, and text modalities are stored in advance. Given a particular bitrate constraint, the adaptation system will select a version having appropriate quality and modality. The content value of the original video is supposed to be 1, and content values of all other content versions are mapped into the range [0,1].

3.1. Obtaining modality curves

3.1.1. Video modality curve

To scale the video, we combine two operations: frame-dropping and requantization. The average PSNR values and bitrates of the scaled video streams are measured to provide the video modality curve. A similar curve can be obtained using the estimation method in [3].

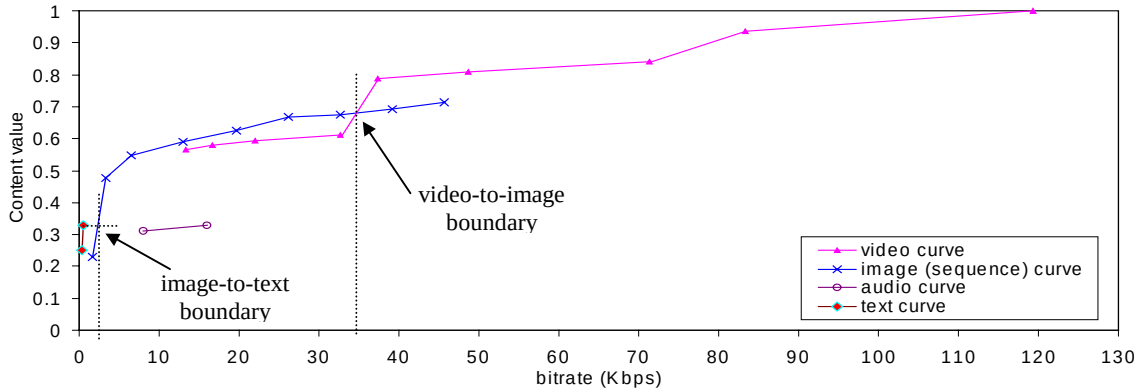


Fig. 3: Overlapped content value model of Foreman video

The content value of video is calculated by multiplying the PSNR values with the scale factor w_1=1/32.81=0.0305 (inverse of maximum PSNR value). This scale factor is used to map the video PSNR values into the content value range [0,1]. The final video modality curve is shown in Fig. 3. From right to left on this curve, the first and second points represent the video streams having no frame-dropping, respectively with quantization factor Q=1 and Q=1.5. The third, fourth and fifth points represent the versions in which all B-frames are dropped, and respectively Q=1, Q=1.5, and Q=2. The sixth to ninth points represent the versions in which all B- and P-frames are dropped, respectively Q=1, Q=1.5, Q=2, and Q=2.5.

3.1.2. Image modality curve

Image sequences are extracted from the full sequence of images (decoded from original video) using the method in [6]. An extracted image sequence is said to represent the "best summary" of the video given the number of images. Thus the scaling operation for image modality is to limit the number of images. Extracted images are encoded in JPEG format such that their qualities are the same as those of the original I-frames. Compared to the full image sequence, any image sequence has an associated "semantic distortion" D, ranging from 0 to infinity [6]. D can be changed into content value as follows:

$$V = w_2 \cdot 1/(1+a \cdot D) \qquad (2)$$

where a is unknown constant and w_2 is scale factor of image modality. We see that, when the image sequence has all frames, the maximum content value of image modality is equal to that of the original video, which was normalized to 1. Thus, w_2 is set to be 1.

It should be noted that (2) is a more general case of the formula V=1/(1+D) proposed in [1]. The constant a, which actually controls the slope of the image modality curve, can be estimated as follows. The video version that contains all original I-frames, called *I-frame stream*, can also be considered as an image sequence, then its content value V^* can be computed from its semantic distortion D^* (provided by the extraction method) as follows:

$$V^* = 1/(1+a \cdot D^*) \qquad (3)$$

Being a video version, the content value of I-frame stream can be evaluated from its PSNR value MV^*_{PSNR}:

$$V^* = w_1 \cdot MV^*_{PSNR} \qquad (4)$$

From (3) and (4) we have:

$$a = \frac{1 - w_1 \cdot MV^*_{PSNR}}{w_1 \cdot MV^*_{PSNR} \cdot D^*} \qquad (5)$$

With the given example, MV^*_{PSNR}=20.01, D^*=119.6, w_1= 0.0305, thus we have a=0.005348. This result gives us the final image curve as shown in the Fig. 3. On the image curve, there are 9 versions of image sequences, with the corresponding numbers of images (from left to right) are 1, 2, 4, 8, 12, 16, 20, 24, and 28 images.

3.1.3. Audio and text modality curves

The MOS scores of audio and text versions are first obtained from subjective tests. These versions are compared to the original video whose equivalent MOS is supposed to be 5. Then users can give the scale factors so as the scaled MOS scores of audio and text show their relevant importance in the range [0,1]. The scale factors of audio and text are found to be 0.1 and 0.07 respectively. Modality curves of audio and text are also depicted in Fig. 3. The content values of audio and text are quite low because for this "performance-like" content, the audio and text modalities cannot describe the semantics as sufficiently as the video or image modalities. Also the audio versions have rather high bitrates, so it has no chance to cut the image curve. That is, the audio versions are never selected. As for the text curve, the bitrates of text versions are very small and close. We find that the bitrates higher than 0.5Kbps are unnecessary for the text modality.

Finally, we can put all modality curves into one model as in Fig. 3. From this model, we can see that the conversion boundary of video-to-image is at 35Kbps and the conversion boundary of image-to-text is at 2.4Kbps.

3.2. Perceptual comparisons of the content versions

From Fig. 3 we note that the bitrates of the I-frame video stream and the extracted 20-image sequence are nearly

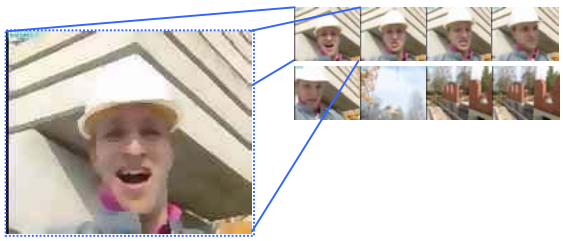
the same (about 33Kbps). However the content value of the video stream is a little lower than that of the image sequence. This is due to the fact that the I-frame stream has fixed intervals between the frames, while the image sequence is extracted based on the semantics of the video content (regardless of B, P, or I-frames), so it would have higher content value. This comparison is shown in Fig. 4. We see that at the scene transition of Foreman video, the extracted image sequence has more information on the panning. However, in the talking scene and construction-site scene, the frames of the two content versions are not much different so they are not illustrated here.

(a) Frames from I-frame video stream

(b) Images from extracted 20-image sequence

Fig. 4: Comparison of the semantics at the transition

(a) Video stream having 20 frames, at 13.3Kbps



(b) Image sequence having 8 images

Fig. 5: Comparison of the visual content of video and image modalities at bitrate 13.3 Kbps.

Fig. 5 shows two other content versions of video and image at the bitrate of 13.3Kbps. This video version still consists of 20 I-frames, but the quantization factor is 2.5. Meanwhile, the image version consists of only 8 images having original spatial quality. We see that, although the image version has fewer images, it still covers enough activity of the video and especially its spatial quality is much better than that of video version. So it is reasonable to select the image version to send to user at this bitrate.

When the bit rate is reduced as low as to 2Kbps for example, the selected output modality will be text. Fig. 6 shows the content comparison of this case. At this time, there are no video versions to be displayed. The image version at this bitrate is a "sequence" of just one image. In this case, the text version would of course give more information than the single image.

Fig. 6: Comparison of image and text versions at 2Kbps

The above experiment shows that using the OCV model in Fig. 3, Foreman video can be converted efficiently to appropriate modalities depending on the bitrate. More details of this case study can be found in [7].

4. CONCLUSIONS

For the purpose of seamless modality conversion, we have presented a systematic approach to help determine the conversion boundaries between modalities. By comparing the content values of different modalities in the overlapped content value model, the adaptation engine can quantitatively make decisions on modality conversion as well as content scaling. Our future works will focus on the efficient estimations of content values across various modalities. The semantics factor will be also explored by considering different genres of multimedia contents.

ACKNOWLEDGMENT

The authors would like to thank Mr. Jae-Gon Kim and Dr. Jeho Nam of ETRI for the fruitful discussions.

REFERENCES

[1] R. Mohan, J. R. Smith, and C.-S. Li, "Adapting multimedia internet content for universal access," *IEEE Trans. Multimedia*, vol. 1, pp. 104-114, Mar. 1999.

[2] ISO/IEC 21000-7 FCD, "Information Technology – Multimedia Framework – Part 7: DIA," July 2003.

[3] J.-G. Kim, Y. Wang, and S.-F. Chang, "Content-adaptive utility based video adaptation," in *Proc. of ICME*, 2003.

[4] A. Vetro, C. Christopoulos, and H. Sun, "An Overview of Video Transcoding Architectures and Techniques," *IEEE Signal Processing Magazine*, pp. 18-29, Mar. 2003.

[5] T. C. Thang et al., "CE Report on Modality Conversion Preference Part-I," ISO/IEC JTC1/SC29/WG11 M9495, Pattaya, Mar. 2003.

[6] H.-C. Lee, S.-D. Kim, "Iterative Key Frame Selection in the Rate-Constraint Environment," *Image Communication*, Issue 18, pp. 1-15, 2003.

[7] T. C. Thang et al., "CE Report on Modality Conversion QoS" ISO/IEC JTC1/SC29/WG11 M10295, Hawaii, Dec. 2003.