Modality conversion for QoS management in universal multimedia access

T.C. Thang, Y.J. Jung and Y.M. Ro

Abstract: Modality conversion 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, modality conversion under the constraint of connection bitrate is considered. Intuitively, when content scaling cannot provide the acceptable QoS, modality conversion may be a good choice to deliver an appropriate quality. From the QoS point of view, two important questions in modality conversion are: 'at what resource constraint point should a change of modality occur?' and 'what is the destination modality?' That is, knowing the conversion boundaries between modalities is crucial for a seamless modality conversion. In this paper, a systematic approach to help answer these questions is presented.

1 Introduction

Universal multimedia access (UMA) is currently a new trend in multimedia communications. A UMA system adapts rich multimedia contents to various constraints of terminals and networks, while providing the best possible quality to the user. In practice, quality of service (QoS) management can be done at both the network level and application level [1]. This paper is concerned with the application level, where content adaptation is an important solution to provide the QoS support. Additionally, the quality in our work is evaluated by a measure that is consistent with human perception; it is not evaluated by traditional physical (objective) measures such as bit error rate or signal-to-noise ratio. In the literature, this kind of quality of service is sometimes called 'quality of experience' or 'perceived quality of service' [2].

Content adaptation has two major aspects: one is modality conversion (also called transmoding), that converts the content from one modality to a different modality; and the other is content scaling, that changes the amount of resource (and so the quality) of the content without converting its modality. Most research on content adaptation has so far dealt with content scaling [3, 4]. However, modality conversion currently appears to be an important issue in UMA [5–7].

The modality concept of multimedia content is actually quite broad. It can be considered from human senses (visual, auditory, tactile, etc.), which have been tackled for a long time in the field of human-computer interfaces (HCI). Modalities can be derived also from different modes of content coding (e.g. video, image, graphics for visual sense). Even different coding formats (e.g. GIF, JPEG) for images are sometimes referred to as modalities or sub-modalities.

There are various conditions that may affect the decision on modality conversion. They can be grouped into four main factors: 1) the modality-presenting capability, which is the support to display certain modalities. This factor can be determined from the characteristics of a terminal (e.g. text-only pager) or the surrounding environment (e.g. a too noisy place); 2) the user preference, which shows the user's levels of interest in different modalities; 3) the resource constraints of terminals or networks, such as the connection bitrate, or the memory size available for the requested contents; 4) the semantics of the content itself. For instance, between a news video and a ballet video, the provider would be more willing to convert the former to a stream of text.

The emergence of MPEG standards, especially MPEG-7 and MPEG-21, facilitates the realisation of UMA systems in an interoperable manner. MPEG-7 [8] defines several classification schemes (CS) to describe various modalities (e.g. ContentCS, GraphicsCodingCS, etc). MPEG-7 also has many tools to describe the semantics (e.g. genres) of multimedia contents. MPEG-21 digital item adaptation (DIA) provides various usage environment description tools to help determine the set of supported modalities, the conversion preference tool to specify user preference on modalities, and the universal constraints description tool to define the (resource) constraints of the adaptation [9].

Currently, modality conversion is mostly carried out when some modality is not supported (e.g. [10]). In this paper, modality conversion is considered mainly in terms of the resource constraint factor. Intuitively, given some resource constraint, the provider may (down)scale the contents to meet the constraint. However, in some cases, the quality of the scaled content 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 video-to-image conversion.

From the QoS point-of-view, the two most important questions for modality conversion are: 'at what resource

IEE Proceedings online no. 20045084

doi: 10.1049/ip-vis:20045084

Paper first received 30th June and in revised form 17th December 2004

The authors are with the Multimedia Group, Information and Communications University (ICU), Yuseong, Daejon, PO Box 77, 305-714, South Korea,

E-mail: yro@icu.ac.kr

constraint point should occur a change of modality?' and, 'what is the destination modality?'

The first question means that at some point, content scaling in the current modality may be no longer effective and another modality will be selected. The second question is clear itself. In this paper, we present a systematic approach to answer these two questions. In order to find the conversion boundaries between modalities, we represent the relationship between the resource and content values (quality) of different modalities using the overlapped content value (OCV) model [5]. However, establishing realistic OCV models is not easy. The main challenge is how to measure the subjective content value when the content is variously scaled and converted. To this end, we present a subjective method to evaluate the content value, which helps us understand the dependency of content value on resource and modalities. As to the composition of the content value, we identify two key quality aspects: the perceptual quality and the semantic quality. The former refers to a user's satisfaction in perceiving the content, regardless of what information the content contains; the latter refers to the amount of information the user obtains from the content, regardless of how the content is presented. Also, we consider computational methods to estimate the content value, which can be used instead of the time-consuming subjective method.

2 Modelling modality conversion for UMA

The process of content scaling can be represented by a 'rate-quality' curve, which shows the quality of a scaled content according to the bitrate (or any resource in general). A recent trend in UMA is to use this rate-quality curve as the metadata to automate content scaling [3, 4]. MPEG-21 DIA provides several description tools (AdaptationQoS) for this type of modelling [9]. Usually, the rate-quality curve is obtained for a particular modality because each modality has its own characteristics. Extending this concept, we introduce the overlapped content value (OCV) model to represent conceptually both content scaling and modality conversion in [5].

An OCV model consists of the rate-quality curves of different modalities (called modality curves) to show the relationship between content values of different modalities. Figure 1 shows an example OCV model of a video content, which consists of video, image, audio, and text curves. The modality curves, provided either manually or automatically, are normally non-decreasing and saturate when the amount of a resource is large enough. We can see that the intersection points of the modality curves represent the conversion boundaries among modalities. Actually, a modality curve is specific to the scaling operation employed (e.g. reducing spatial/temporal resolutions, requantising, or any combination of these). That is, there may be multiple operation curves for each modality, corresponding to

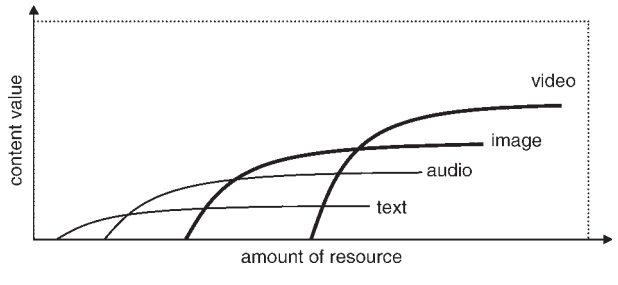

Fig. 1 *Overlapped content value model of video content*

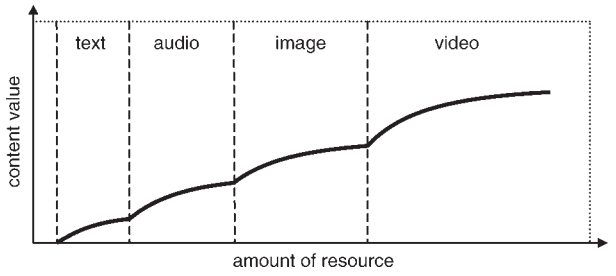

Fig. 2 *The final content value function of video content*

different scaling operations. For simplicity, just one curve is drawn for each modality, representing the scaling operation selected for the session.

Let $VM_j(R)$ denote the rate-quality curve of modality j of a content, $j = 1 \ldots J$, where J is the number of modalities of the content; R is the amount of resource. $VM_j(R) \geq 0 \forall j$. Also let w_j denote the scale factor of modality j. The content value function, which is the convex hull of the modality curves in the OCV model, can be written as follows:

$$V = \max\{w_j \times VM_j(R) | j = 1 \ldots J\} \quad (1)$$

By the proper estimation of content value for different modalities, we can put the modality curves into an OCV model. When the content values of some modalities are measured elsewhere using some different scoring scales (e.g. from 0 to 10 or from 0 to 100), the scale factors can be used to map the content values into a common scale. Figure 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 in addition to content scaling, to maintain an acceptable quality. As mentioned, building the OCV model is a not a simple task owing to the challenge of quantifying the subjective content value. In the following Section, we study the evaluation of content value within the context of content adaptation.

3 Content value evaluation

3.1 Aspects of content value

It is commonly agreed that multimedia quality has a multidimensional nature and that key quality dimensions/aspects should be identified for the application in use [11]. With video content scaling, the quality is normally evaluated by some measures that show the perceptual satisfaction of scaled video. In the case of video-to-image conversion, some semantic scores, representing the understandability of the key-frame set, are often mentioned [12, 13]. In some extreme cases, dance video converted to text for example, it is obviously not the perceptual quality, but the amount of conveyed semantics that really counts.

For content adaptation, we contend that the content value consists of both the perceptual quality (PQ) and the semantic quality (SQ). Although one may say that PQ already includes SQ, the separation is necessary because when PQ is reduced (e.g. lower frame rate), SQ may remain unchanged [14]. Even when PQ is nearly zero (e.g. the above video-to-text conversion), the value of SQ may still be acceptable. We then propose the composition of content value as follows:

$$V = s \times PQ + (1 - s) \times SQ \quad (2)$$

where s is the weight of PQ, $0 \leq s \leq 1$; s can be assigned by the provider depending on the particular applications.

In general, if the information itself is more important than the perceptual satisfaction, then the weight of *SQ* should be higher than that of *PQ*. We assume that an average user in a normal situation would need *SQ* and *PQ* in an equal manner. So, as default, we let the value of *s* be 0.5. In the following Sections, the evaluations of *SQ* and *PQ* will be presented.

3.2 Subjective method

The quality can be represented by some 'physical' measures, e.g. PSNR. However, users are the ultimate judges of the quality. In this Section, we present a practical procedure to evaluate the two above quality measures of the adapted contents.

Owing to the fact that the content value depends on many quality dimensions, we should clearly instruct the subjects so that every subject pays similar attention to the contents, thus guaranteeing a stable evaluation.

We are interested in the quality (both semantic and perceptual) of the adapted (or test) version of the content compared to the original one. Therefore, each time during the test, the subject should be presented with two content versions, the original and then the adapted one, so that the subject can give the score to the adapted version with respect to the original one. This feature is similar to the degradation category rating (DCR) method specified in ITU-T Rec. P.910 [15].

As for the measures of quality, with every adapted version, we ask the subjects to give two scores, one for the 'understanding' (i.e. semantic quality) and one for the perceptual quality. The understanding score is explained as the perceived amount of information conveyed by the version, regardless of how the version is presented, while the perceptual quality is defined as the satisfaction of the subject while perceiving the version, regardless of what information is conveyed. Each score will take an integer value in a Likert-style ten-point scale, from 0 to 9 [16]. For 'understanding', a score of 9 means that the adapted version shows sufficiently the original semantics, whereas a score of 0 means that the adapted version has totally different semantics compared to the original. For perceptual quality, a score of 9 means that the adapted version has the same presentation quality as the original, while a score of 0 means a very annoying and/or totally different presentation.

On the scoring scale, we only have the explanations at the two ends (0 and 9). There are no descriptions for the intermediate levels because such labels could not be conceptually equal and may even mislead the subjects [11]. The score range from 0 to 9 is selected because with modality conversion and content scaling, the quality levels can vary widely, and it is easier for people to evaluate the quality using a ten-point scoring scale.

The final score for each content test version is the mean score of all subjects. The inter-subject reliability of the test is checked using confidence intervals as specified in ITU-R Rec. BT.500 [17].

The subjective evaluation is of high importance because it can be applied to various cases of different modalities and contents. It is the key tool to obtain the OCV model in our work. However, subjective tests are expensive and time-consuming, so computational methods used to estimate the quality have long been an interesting research topic.

3.3 Computational methods

It is expected that the computational evaluations for various modalities would be very different. In this Section, we focus on video and image modalities that are the most popular in practice. It should be noted that, the actual object of the adaptation in this part is a video shot, which has been segmented in advance.

3.3.1 Computational estimation of perceptual quality:

PSNR has been a popular objective quality measure for a long time. In [7], we use PSNR for the *PQ* of video modality and implicitly suppose that it is representative of the content value. However, the problem is that the PSNR measure is not well correlated to human evaluation [18–21], especially with the case of digital video where blockiness and blurriness artifacts are very common. The human evaluation of perceptual quality is normally measured by the mean opinion score (MOS), which is obtained by subjective tests [15]. Recently, there has been a significant amount of work that deals with estimating the 'objective MOS', where the human judgment on quality is modelled by exploiting knowledge about the human visual system (HVS). An estimation method of objective MOS is referenced [18, 19] if the method compares the adapted version with the original one, and non-referenced [20, 21] if the method just considers the artifacts in the adapted version without comparing to the original. Current objective MOSs are shown to be much more consistent with subjective evaluation than the PSNR measure [18–21].

In this paper, the objective MOS proposed in [18, 19] is used instead of PSNR to measure the *PQ* of video and image modalities. This method is referenced, thus it is suitable for obtaining the quality of the adapted version with respect to the original. The basic procedure of this estimation method is as follows. First, some quality features, which are significant to human perception (e.g. edges, contrast, temporal activity, etc.), are extracted and enhanced by some perceptual filters. The quality features of the adapted version are compared to those of the original version to obtain a set of quality parameters that are indicative of perceptual quality changes. These quality parameters are then used to 'deduce' a quality score of the adapted version using some quality models that emulate the HVS functions (e.g. visual masking, error pooling, etc.) [18]. We find that the quality model that incorporates both temporal and spatial quality features has the highest accuracy. The reason is that, in our experiment, the video is scaled widely in both the temporal and spatial domains. The formula used to compute the estimated distortion is given as follows [19]:

$$\begin{aligned} VQM = &- 0.2097 \times si\_loss \\ &+ 0.5969 \times hv\_loss \\ &+ 0.2483 \times hv\_gain \\ &+ 0.0192 \times chroma\_spread \\ &- 2.3416 \times si\_gain \\ &+ 0.0431 \times ct\_ati\_gain \\ &+ 0.0076 \times chroma\_extreme \end{aligned} \tag{3}$$

where *VQM* (video quality metric) is the estimated distortion, $0 \leq VQM \leq 1$; *si_loss* is the quality parameter that detects a decrease of spatial information; *hv_loss* is the quality parameter that detects a shift of edges from horizontal and vertical orientations to diagonal orientations; *hv_gain* detects a shift of edges from diagonal orientations to horizontal and vertical orientations; *chroma_spread* detects changes in the colour sample distribution; *si_gain* detects the quality improvements that result from edge sharpening or enhancements; *ct_ati_gain* is the product of the contrast feature and the temporal information feature; and *chroma_extreme* detects severe localised colour impairments [19].

The detailed descriptions and computations of these quality parameters can be found in [18, 19].

Then, the value of PQ, measured on the 0-9 scoring scale, is mapped from VQM as follows:

$$PQ = 9 \times (1 - VQM) \tag{4}$$

3.3.2 Computational estimation of semantic quality: In contrast to PQ, the computation of SQ has been studied very little in the literature We denote SQ_{image} and SQ_{video} the semantic qualities of image and video modalities. Without loss of generality, in this Section, these qualities are normalised in the range [0,1]. It should be again noted that we are interested in the quality of an adapted version compared to the original version, not to a different content (e.g. a sport video compared to a music video).

For video-to-image conversion, some key-frame extraction methods can be applied in order to get a sequence of 'important' images from the original video (shot). Each image has the same quality as the corresponding frame in the original video. The images are said to be 'important' because they are selected based on some semantically salient features, e.g. motion activity, colour, etc. The extraction methods assign to each image sequence (i.e. a set of key-frames) a 'semantic distortion' D, ranging from 0 to infinity [12, 13]. Each extraction method has its own way to compute D. The extraction method of [12], which will be used in our experiment, computes D as follows. First, a semantically salient content feature vector showing the instant temporal activity is computed at each timestamp (i.e. for each frame in the original video). There is a wide range of possible content features, such as the intensity variance or the histogram of each frame [12]. In our experiments, the MPEG-7 scalable colour descriptor is used as the content feature. Then, D is defined as the sum of the content feature differences, taken over all timestamps, between a given image sequence and the full sequence (i.e. the original video). Here, for an image sequence, the key-frames are repeated to fill the 'empty' timestamps. In principle, among the image sequences of the same number of images, the sequence having the smallest D is the extracted (or selected) sequence.

SQ_{image} is actually the semantic quality of an image sequence (also called an image version) compared to the original video. We represent SQ_{image} according to D as follows:

$$SQ_{image} = \frac{1}{1 + a \times D} \tag{5}$$

where a is a constant that controls the slope of the function; a depends on content characteristics and on the way D is computed.

As to the semantic quality of video modality, in [22], the authors propose an importance value for a video object based on the product of motion activity and spatial complexity. In [23], the utility of a video shot is defined as the product of the shot's spatial complexity and duration. Generally, the SQ_{video} of an adapted video version, compared to the original version, is composed of temporal semantic quality ($SQ^{temporal}_{video}$) and spatial semantic quality ($SQ^{spatial}_{video}$), which are respectively affected by temporal and spatial content scaling operations [Note 1]. For example, $SQ^{temporal}_{video}$ will be reduced if some frames are dropped

Note 1: In our work, content scaling in the spatial domain is limited to requantisation, i.e. excluding spatial resolution scaling.

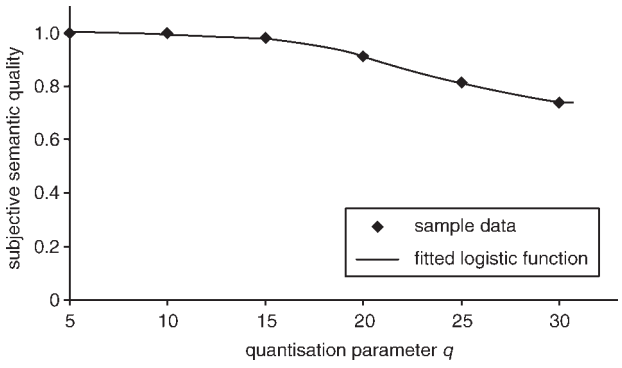

Fig. 3 *Example of relationship between subjective $SQ^{spatial}_{video}$ and q for the foreman video*

The sample data show 6 pairs of ($SQ^{spatial}_{video}$, q), where q takes the values of 5, 10, 15, 20, 25, and 30. The curve is the logistic function (7) fitted to the sample data

(i.e. scaling in the temporal domain), and $SQ^{spatial}_{video}$ will be reduced if the quantisation parameter is increased (i.e. scaling in the spatial domain). In an extreme case, when the degradation in the spatial domain is too severe, e.g. all spatial details are lost, both $SQ^{spatial}_{video}$ and SQ_{video} would approach zero. Similarly, when the degradation in the temporal domain is too severe, $SQ^{temporal}_{video}$ and SQ_{video} would also approach zero. So, SQ_{video} can be combined from $SQ^{temporal}_{video}$ and $SQ^{spatial}_{video}$ as follows:

$$SQ_{video} = SQ^{temporal}_{video} \times SQ^{spatial}_{video} \tag{6}$$

$SQ^{temporal}_{video}$ is reduced if frame dropping is used for content scaling, so similar to SQ_{image}, the value of $SQ^{temporal}_{video}$ can be represented by (5). Here, the frames to be dropped are not determined by the image extraction method, but by the frame dropping policy of a video transcoder (e.g. dropping all B frames, or dropping all B and P frames); the computation of D is still the same.

Alternatively, if only requantisation is applied for content scaling in the spatial domain, we can assume that $SQ^{spatial}_{video}$ is affected mainly by the quantisation parameter q. Generally, the relationship of $SQ^{spatial}_{video}$ and q has the S-shape [24], where the quality usually reaches its maximum value when q approaches 1 and its minimum value when q approaches 31. Figure 3 shows an example of the relationship between $SQ^{spatial}_{video}$ and q for the 'foreman' video encoded in MPEG-4 format (simple profile). Note that in this Figure, the value of q is varied but the original frame rate is fixed. Basically, the S-shape can be represented by different analytical forms, of which one possibility is the well-known logistic function [17]:

$$SQ^{spatial}_{video} = b + \frac{1 - b}{1 + e^{c \times (q - d)}} \tag{7}$$

where b is the minimum value of $SQ^{spatial}_{video}$, $b \leq SQ^{spatial}_{video} \leq 1$, $0 \leq b \leq 1$; the flexion point of the curve is at $q = d$; and c controls the slope of the curve.

In contrast to these analytical model-based methods, the utility estimation in [4] tries to classify the video shots, using some content features, into a number of classes. Each class has a regression model mapping content features to utility, which can be obtained by a machine learning approach. This method is general for different application domains, but it is complex and has been checked with the PSNR measure only. In our experiments, the analytical model-based methods will be employed.

4 Experiments

In this Section, we will explore the possibilities of modality conversion for some streaming contents The operations of content scaling and modality conversion in our experiment are carried out offline. For a given original content, a number of adapted versions of different modalities are produced and stored in advance. Given a bitrate constraint, the adaptation system will select a version having appropriate quality and modality to send to the user. For this purpose, the OCV models of the content, consisting of different modality curves, are first obtained using the subjective method. The obtained models are used to help answer the two basic questions raised in Section 1. Then, a case study, in which we apply the computational methods to estimate the content value of video and image modalities, is considered.

4.1 Subjective experiments

4.1.1 Experimental setup: We have three original contents. The first content is a landscape video consisting of 240 frames (without audio channel) extracted from the Lascaux stream of the MPEG-7 database. The second content is the foreman video (without audio channel) consisting of 300 frames. These two contents, originally encoded in MPEG-4 format, have a luminance frame size of 176×144 (QCIF), quantisation parameter $q = 10$, GOP structure of $M = 3$ and $N = 15$, and a frame rate of 25 fps. The third content, an audiovisual (i.e. 'audio-video', denoted as AV) clip extracted from the eye-exam stream of the MPEG-7 database, is educational content consisting of both audio and video channels. The video channel, consisting of 330 frames, is encoded in the same manner as the above content, except that its luminance frame size is 256×174. The original audio channel is encoded at 24kbits using the XingMPEG® encoder [25]. Some sample frames extracted from these three contents are shown in Fig. 4.

For the first and second contents, to obtain the adapted video versions, the original videos are scaled using a combination of frame-dropping and requantisation. Image sequences are obtained as the key-frames of the original video using the method in [12]. The scaling operation for the image modality is essentially to limit the number of images. Extracted images are encoded in the JPEG format such that their qualities are the same as the I-frames of the original video. Text, which is the description of the original content, is created manually. There is only one text version because its bitrate is very small. Audio content is created as the speech from the text. Then the XingMPEG® encoder is used to produce two audio versions having bitrates of 24kbits and 8kbits respectively (using MPEG-2 audio layer 2 format). All test versions of these two contents, together with their modalities and bitrates, are listed in Table 1.

As for the third content, the adapted AV versions are produced by scaling the video channel in the same way as the previous two examples. Then, image sequences are extracted and combined with the original audio channel to create the 'audio-image' (AI) versions. That is, the audio channel is kept intact in AV and AI versions. The audio versions are obtained from the audio channel of the original content. And a text version is produced as the script of the teacher's voice in the audio channel. All test versions of the educational content are listed in Table 2.

The test versions are presented on a 20^{Prime} Apple Cinema LCD Monitor, at a resolution of 1280×768 and with progressive display. The colour of the monitor background is set to 50% grey.

Eighteen non-expert subjects were recruited to participate in the experiments. All subjects have normal colour vision, normal visual acuity or wear corrective glasses. The tests are carried out based on the procedure presented in Section 3.2. The instructions and explanation of the quality scores are provided in written form. The subjects are asked to pay equal attention to the temporal and spatial domains. The semantics of the original content includes both the spatial details and the temporal changes in the scene.

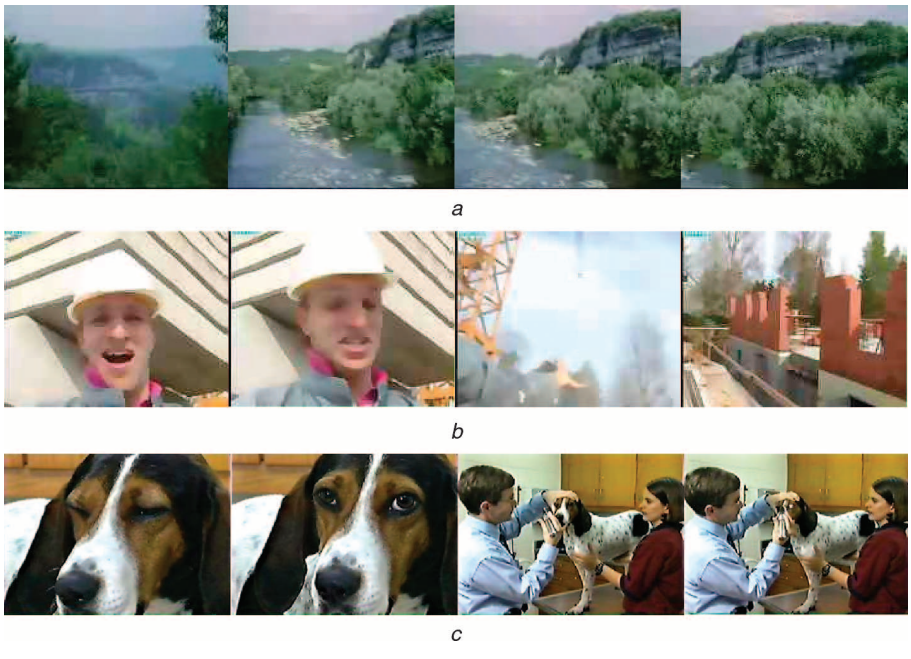

Fig. 4 *Extracted sample images*
a Landscape content
b Foreman content
c Educational content

Table 1: List of test versions and their characteristics for the landscape and foreman contents

No.	Modality	Bitrate (kbit/s) Landscape	Foreman	Description
1	Video	80.00	119.33	Original video, $q = 10$, $f = 25$ fps
2	Video	45.78	71.33	Dropping all B frames, $q = 10$, $f = 8.3$ fps
3	Video	26.63	32.67	Dropping all B and P frames, $q = 10$, $f = 1.7$ fps
4	Video	13.36	26.33	Dropping all B frames, $q = 30$, $f = 8.3$ fps
5	Video	7.5	11.33	Dropping all B and P frames, $q = 30$, $f = 1.7$ fps
6	Image	53.33	52.27	Sequence of 32 images, $q = 10$
7	Image	26.67	26.14	Sequence of 16 images, $q = 10$
8	Image	13.33	13.07	Sequence of 8 images, $q = 10$
9	Image	6.67	6.53	Sequence of 4 images, $q = 10$
10	Image	1.66	1.63	Sequence of 1 image, $q = 10$
11	Text	0.5	0.5	A stream of explanatory text
12	Audio	24	24	Explanatory speech with high quality
13	Audio	8	8	Explanatory speech with low quality

The first version is the original. The other versions are adapted versions, each has a certain modality, bitrate, and coding parameters. Note: f is the frame rate

Table 2: List of test versions and their characteristics for the educational content

No.	Modality	Bitrate (kbps)	Description
1	AV	264.73	AV with original audio and video, $q = 10$, $f = 25$ fps
2	AV	157.82	AV with scaled video (dropping all B frames, $q = 10$, $f = 8.3$ fps)
3	AV	91.64	AV with scaled video (dropping all B and P frames, $q = 10$, $f = 1.7$ fps)
4	AV	72.73	AV with scaled video (drop all B frames, $q = 30$, $f = 8.3$ fps)
5	AV	48.73	AV with scaled video (dropping all B and P frames, $q = 30$, $f = 1.7$ fps)
6	AI	124.07	AI with scaled image sequence (32 images, $q = 10$)
7	AI	74.04	AI with scaled image sequence (16 images, $q = 10$)
8	AI	49.02	AI with scaled image sequence (8 images, $q = 10$)
9	AI	36.51	AI with scaled image sequence (4 images, $q = 10$)
10	AI	27.13	AI with scaled image sequence (1 image, $q = 10$)
11	Audio	24	Original audio channel
12	Audio	8	Scaled audio with low quality
13	Text	0.5	Textual script of the speech in audio channel

The first version is the original. The other versions are adapted versions, each has a certain modality, bitrate, and coding parameters. Note: AV means 'audio-video', AI means 'audio-image', and f is the frame rate

Before testing, some examples are displayed to the subjects so they can understand the adaptation range. During the real test, the test versions are shown randomly, so the subjects are not biased by *a priori* knowledge of presentation ordering.

4.1.2 Results and analysis: The screening of subjects is carried out according to ITU-R Rec. BT.500 [17]. The result is that no subjects have been rejected. Then, the mean quality scores and the associated 95% confidence intervals are calculated. Figure 5 shows the perceptual and semantic quality curves for different modalities of the contents. The final OCV models, obtained by averaging *PQ* and *SQ* (i.e. $s = 0.5$ for all modalities), are shown in Fig. 6. Note that the maximum content value is 9.

We see that the relationship of *SQ* and *PQ*, in which *SQ* is always higher than or equal to *PQ*, is quite consistent (Fig. 5). This is perhaps due to the fact that the users use their experience in reasoning and understanding; a small clue of information in a degraded presentation may be enough for the users' understanding. Especially, the *PQ* of text modality is nearly zero; however, its *SQ* is quite significant. That is, in a critical application (e.g. very low bandwidth communication) where the semantics is the most important thing, the conversion to text is really helpful. Although having the same information, the good audio version has a little higher *SQ* and *PQ* than the text version. The reason is that, for users to perceive and catch the content, listening is more comfortable than reading.

The 95% confidence intervals of quality scores for different modalities (averaged for both landscape and foreman contents) are given in Fig. 7. The overall average value of the confidence intervals is ± 0.473 on the 0–9 scale (equivalent to ± 5.26 on the 0–100 scale). This result is good compared to other tests for video streaming over the Internet [20] and wireless networks [21], in which the average confidence interval is from ± 7.8 to ± 8.5 on the 0–100 scale. With the educational content, the overall average value of the confidence intervals is ± 0.433 on the 0–9 scale. The results of confidence intervals show a good agreement between subjects. That means the tests are reliable, and also *PQ* and *SQ* are all meaningful to the users. However, as seen in Fig. 7, *SQ* usually has a larger confidence interval than *PQ*, that means giving semantic scores is more difficult than giving perceptual scores. We also see that the confidence intervals (i.e. score variances) corresponding to converted modalities are usually higher than those of the original modality, except the case of the *PQ* of text and audio where subjects usually give a score of 0 or 1 on the 0–9 scoring scale. This finding implies that the increased variations of

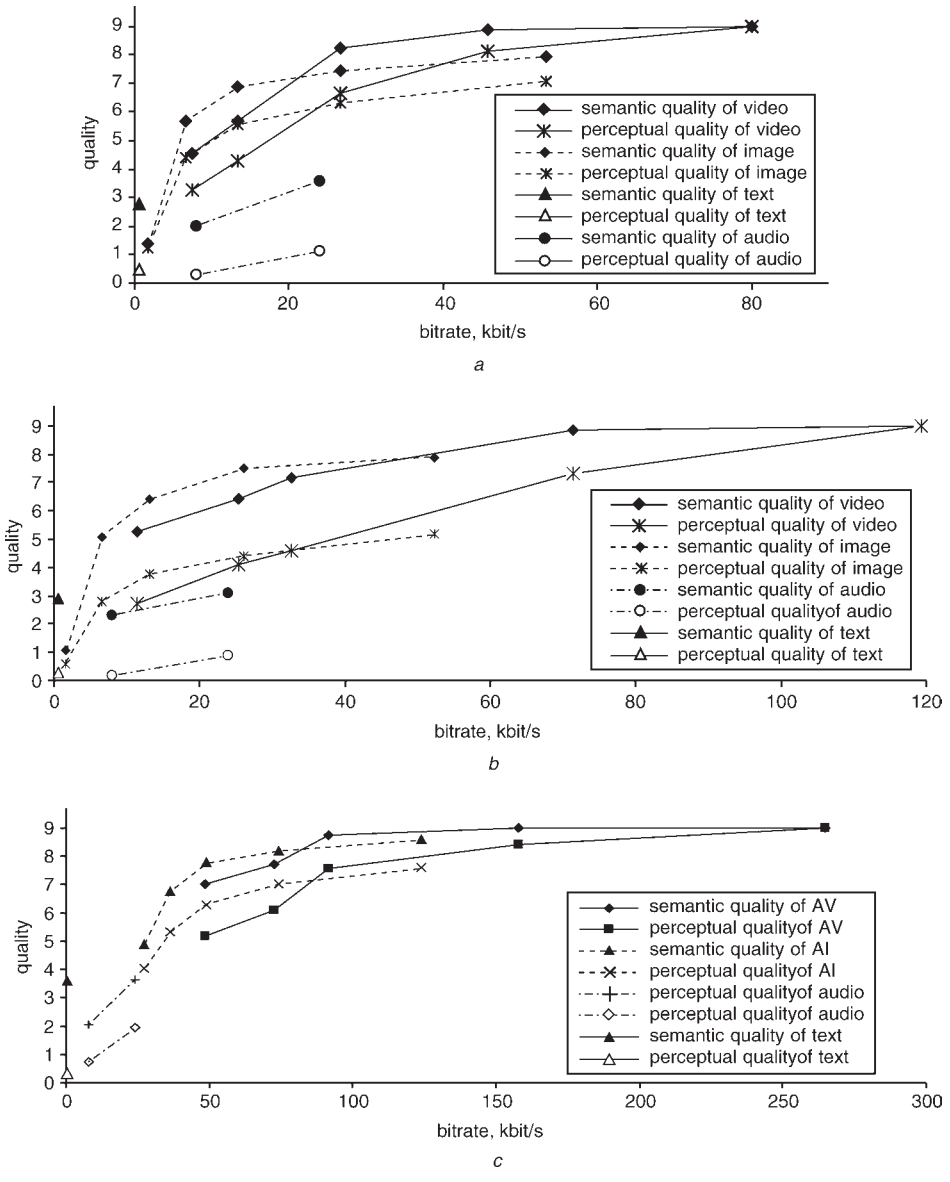

Fig. 5 *Quality curves of different modalities for the three contents*

a Landscape
b Foreman
c Educational

subjective quality owing to modality conversion should be considered in content adaptation.

From the final OCV model of the landscape content (Fig. 6*a*), we find that the conversion point between video and image modalities is at 23 kbits and the conversion point between image and text is at 1.7 kbits. There is no conversion point related to audio modality because its content value is similar to that of text while its bitrate is rather high. The obtained OCV model confirms that modality conversion, specifically video-to-image and image-to-text, is useful for this video content. A similar phenomenon can be found with the foreman content (Fig. 6*b*). The difference between these two contents is that the foreman video has a little higher motion activity, then the bitrates of scaled video versions of the foreman video are higher than those of the landscape video. And thus the conversion point of video-to-image is rather high, at about 32 kbits.

As for the educational content (Fig. 6*c*), we see that, depending on the bitrate constraint, the original AV modality can be converted to either AI, audio, or text modalities. Also we see that the combination of an audio channel with either a low quality video channel or a sequence of several images (even just a single image) gives a rather good content value, compared to the audio modality. This is actually a phenomenon of the cross-modal influence, resulting in a synergy between the element modalities, which makes the quality of a combined modality (e.g. audiovisual) much higher than that of an element modality (e.g. audio only or video only) [26].

4.2 Case study using computational methods

Now we use the computational methods to obtain the perceptual quality and semantic quality for video and image modalities. The landscape and foreman videos are again used in this Section.

The *PQ* of scaled video versions, compared to the original version, is obtained as the objective MOS using (4). This objective MOS has been verified extensively and high reliability has been reported [18, 19]. The *PQ* of image sequences is obtained in the same way as a video version. Yet, image sequences need preprocessing such that they can be treated as a video. Specifically, key-frames of an image

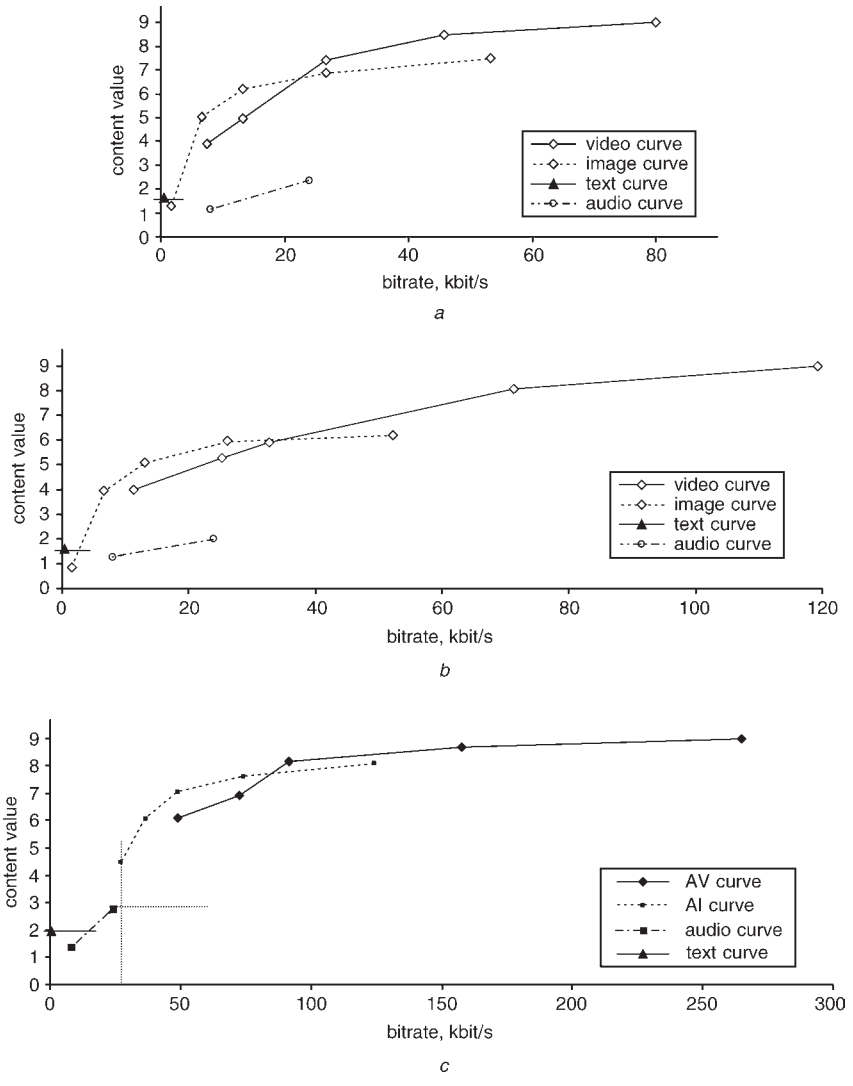

Fig. 6 *Final modality curves in OCV models for the three contents*

$s = 0.5$ for all modalities
a Landscape
b Foreman
c Educational

sequence are repeated to build a new sequence, of which the 'frame-rate' is the same as the original video. This new sequence is then compared to the original video to get the objective MOS.

The *SQ* of image modality is computed using (5). Image sequences (i.e. image versions) are extracted from the original video using the method in [12]. The parameter *a* of each content is estimated by fitting the function, using the mean square error criterion, to some empirical training data. For the landscape video, the

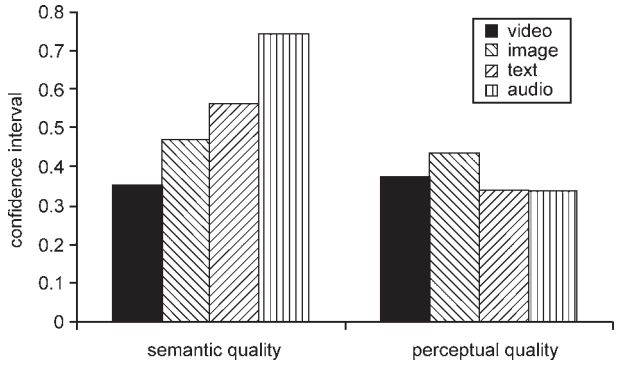

Fig. 7 *Confidence intervals of perceptual and semantic qualities with different modalities*

The results are averaged for the landscape and foreman videos. The overall average confidence interval is ±0.473 on the 0–9 scale

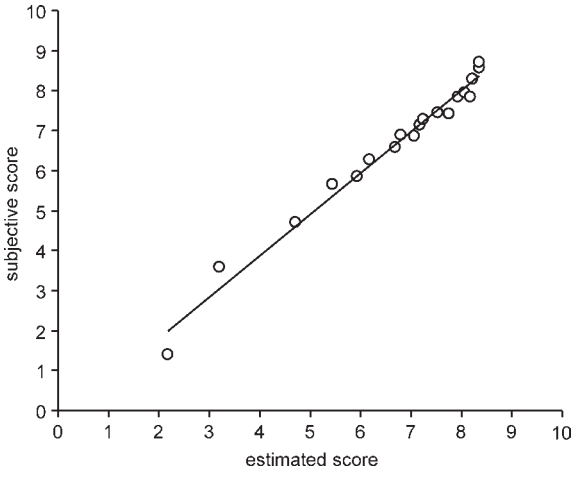

Fig. 8 *Semantic qualities of image modality for the landscape content: estimated scores against subjective scores*

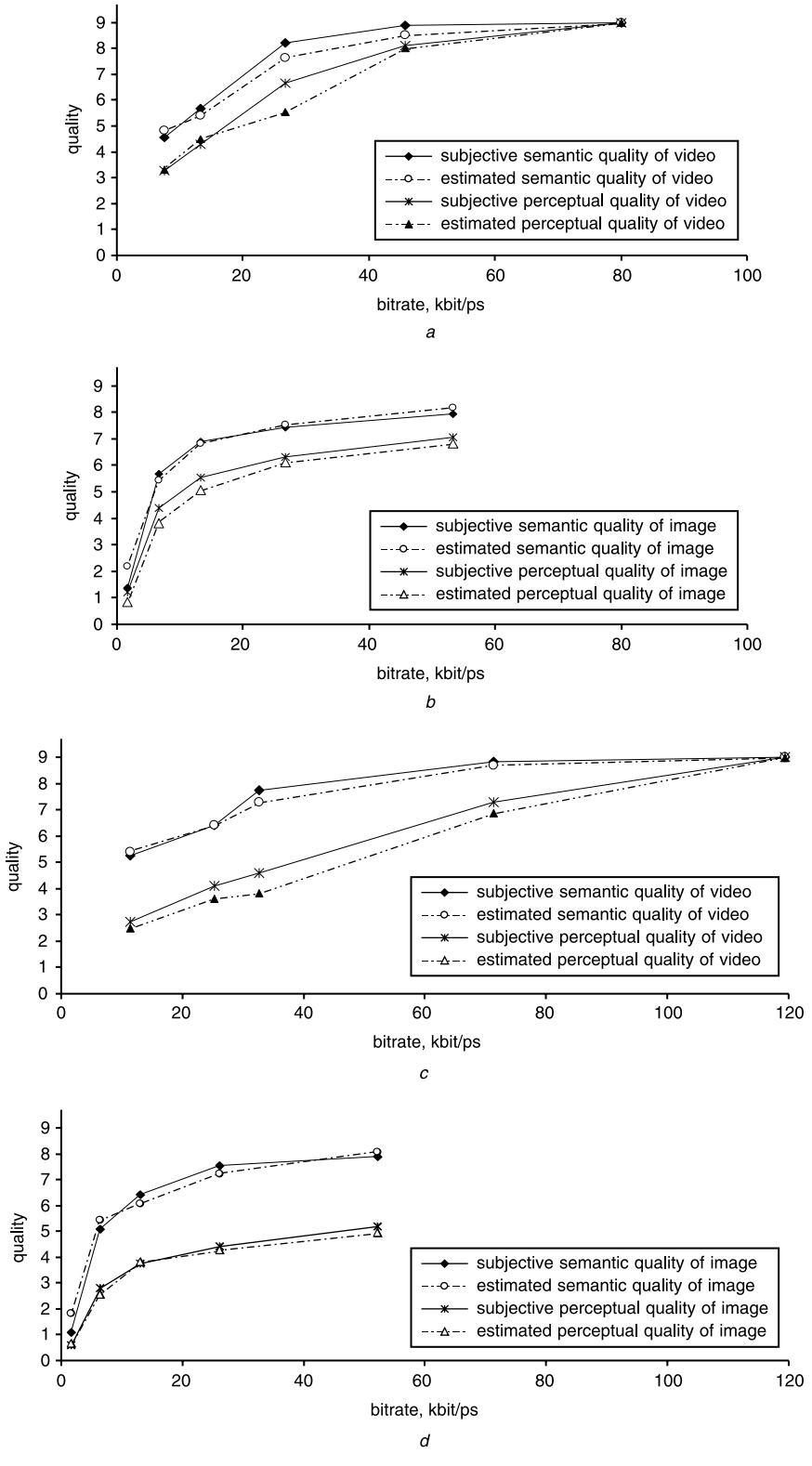

Fig. 9 *Comparison of subjective and estimated qualities for landscape content and Foreman content. The quality curves of video and image modalities are separated for the purpose of clarity*

a, *b* Landscape content
c, *d* Foreman content

training data are five (D, SQ_{image}) pairs corresponding to the five image versions listed in Table 1. The obtained value of a is 0.0023. In Fig. 8, the estimated SQ_{image} values are checked against subjective SQ_{image} values for a set of 20 image sequences extracted from the landscape video. The numbers of images of these sequences take the even integer values in the range from 2 to 48, except the values 4, 8, 16, and 32 (corresponding to the sequences of 4, 8, 16, and 32 images that are used in the subjective test). The subjective SQ_{image} values of these 20 image sequences are obtained from a separate subjective test. The resulting Pearson correlation between the estimated scores and the subjective scores is 0.97. Note that the SQ obtained by (5) has a maximum value of 1, thus we need to rescale this quality into the 0–9 scoring scale of the subjective test using the scale factor of 9. Similarly, with the foreman video, the value of a is 0.0063 and the corresponding Pearson correlation is 0.98.

The SQ of video modality is computed using (6). First, $SQ_{video}^{temporal}$ is estimated using (5) as the above (i.e. using the same a for both video and image modalities), with D computed for the frame-dropped video versions (specifically, all B frame dropped and all B and P frame dropped versions). Second, $SQ_{video}^{spatial}$ is computed using the logistic function (7). The curve in Fig. 3 shows an example of the logistic function fitted to the subjective data of the foreman video, where $b = 0.72$, $c = 0.33$, and $d = 22.6$. The overall SQ of video modality is then computed as the product of (5) and (7). Suppose that to scale the video content, we can vary q with 5 levels (namely 10, 15, 20, 25, 30) and change the frame-dropping with 3 levels (namely no-dropping, B frame dropping, B and P frame dropping). Then we have totally 15 video versions, which is, in our experience, enough for selecting an appropriate video version in practical cases. The Pearson correlation of the estimated SQ_{video} with the subjective SQ_{video} is calculated over the set of these 15 video versions, and the resulting correlation is 0.95. For the landscape content, the corresponding values are $b = 0.64$, $c = 0.31$, $d = 20.9$, and the Pearson correlation is 0.96. These correlation results mean the estimated SQ_{video} is suitable to replace the subjective SQ_{video}.

Finally, the estimated quality curves of the landscape and foreman videos are shown together with the subjective counterparts in Fig. 9. For the purpose of clear comparison, the video quality curves and image quality curves are depicted separately. We see that the estimated qualities are consistent with the subjective qualities.

The above experiments show that modality conversion is a good choice to widen the range of QoS control for UMA. Furthermore, the subjective method and computational methods to evaluate the content value can effectively help to build realistic OCV models, which are used to make decisions on modality conversion.

It should be noted that, similar to the rate-quality curves [3, 4], the OCV model is applicable to both offline and online adaptations. That is, the model can describe not only some pre-transcoded versions (offline case) but also the potential scaling operations for a content (online case). In the online case, approximation can be used to estimate the content value of potential content versions. For example, the formulation of the semantic quality of image modality is built analytically based on 5 sample points (5 image versions), but this formulation can be used to estimate the quality of other image versions, including the potential versions.

5 Related work

In addition to the constraints of modality-presenting capability and resource, modality conversion can be studied from some other perspectives. In [5], we show how the user preference on conversion can be specified efficiently and then, based on a conceptual OCV model, we propose a way to incorporate the preference into the adaptation process. In the context of HCI, the modality selection (conversion) is also studied with respect to different device types [27]. In one sense, this can be considered as one case of modality conversion according to the user preference, where users have different preferences when using different devices. Additionally, semantics analysis would give many useful hints to modality conversion [28]. The cross-modal influence on the overall (e.g. audiovisual) quality, as well as on individual channel (e.g. audio or visual) quality, is also an important semantics-related issue (e.g. [26]). However, research on this issue is still at the initial stage. Currently, there is also some related work that employs metadata to automate the conversion process between specific modalities/formats [29, 30].

There is also some work studying multidimensional subjective quality. In [14, 31], video quality is evaluated with different frame rates. The users' satisfaction (i.e. PQ) is obtained subjectively, whereas the informational transfer (i.e. SQ) is obtained 'objectively' by counting the users' correct answers to a predefined question list that concerns the semantics of a video. The work in [32] deals with football content, where video and animation presentations are studied, focusing on the reality and enjoyment compared to the real game. Our work is different in that we consider the 'fidelity' of the adapted content with respect to the original and in that the adapted content is varied in terms of both quality and modality. Moreover, SQ in our work is not computed based on 'correct answers' as in [14, 31, 32] because the questions are specific for a given content and may not sufficiently cover the semantics of the content.

6 Conclusions

For the purpose of seamless modality conversion, we have presented a systematic approach to help determine the conversion boundaries between modalities. We presented the overlapped content value (OCV) model to represent the dependence of content value on resource and modalities. In the context of content adaptation, we pointed out two important aspects of content value: the perceptual quality and the semantic quality. Then we presented the subjective method to evaluate the content value of a content that may be drastically scaled or converted to different modalities. For the specific case of video and image modalities, we discussed some computational methods to replace the time-consuming subjective evaluation. Finally, by comparing the content values of different modalities in the obtained OCV model, the adaptation engine can quantitatively make decisions on modality conversion in addition to content scaling. Our future work will focus on the objective evaluation of content value across other modalities (e.g. graphics, 3-D video). The semantics factor will be also explored by considering some genre hierarchy of multimedia contents.

7 References

1 Chandra, S., Ellis, C.S., and Vahdat, A.: 'Application-level differentiated multimedia Web services using quality aware transcoding', *IEEE J. Sel. Areas Commun.*, 2000, **18**, (12), pp. 2544–2565
2 Pereira, F.: 'Content adaptation: the panacea for usage diversity?, in Visual content processing and representation'. 'Lectures Notes in Computer Science' (Springer, New York, 2003), pp. 9–12
3 Vetro, A., Christopoulos, C., and Sun, H.: 'Video transcoding architectures and techniques: an overview', *IEEE Signal Process. Mag.*, 2003, **20**, (2), pp. 18–29
4 Wang, Y., Kim, J.-G., and Chang, S.-F.: 'Content-based utility function prediction for real-time MPEG-4 video transcoding'. Proc. Int. Conf. on Image Processing, 2003, Vol. 1, pp. 189–192
5 Thang, T.C., Jung, Y.J., and Ro, Y.M.: 'Modality conversion in content adaptation for universal multimedia access'. Proc. Int. Conf. Imaging Science, Systems, and Technology, Nevada, USA, 2003, pp. 434–440
6 Vetro, A.: 'MPEG-21 Digital Item Adaptation: enabling Universal Multimedia Access', *IEEE Multimedia*, 2004, **11**, (1), pp. 84–87
7 Thang, T.C., Jung, Y.J., Lee, J.W., and Ro, Y.M.: 'Modality conversion for universal multimedia services'. Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services, Lisboa, 2004
8 ISO/IEC IS 15938-5:2001: 'Information technology - multimedia content description interface - multimedia description schemes', 2003
9 ISO/IEC FDIS 21000-7: 'Information technology - multimedia framework - Part 7: DIA', 2003
10 Kaup, A.: 'Video analysis for universal multimedia messaging'. Proc. 5th IEEE Southwest Symp. Image Analysis and Interpretation, 2002, pp. 211–215
11 Watson, A., and Sasse, M.A.: 'Measuring perceived quality of speech and video in multimedia conferencing applications'. Proc. ACM Multimedia 98, Bristol, UK, 1998, pp. 55–60

12 Lee, H.-C., and Kim, S.-D.: 'Iterative key frame selection in the rate-constraint environment', *Signal Process., Image Commun.*, 2003, **18**, (1), pp. 1–15
13 Chang, H.S., Sull, S., and Lee, S.U.: 'Efficient video indexing scheme for content-based retrieval', *IEEE Trans. Circuits Syst. Video Technol.*, 1999, **9**, (8), pp. 1269–1279
14 Ghinea, G., and Thomas, J.P.: 'QoS impact on user perception and understanding of multimedia clips'. Proc. ACM Multimedia 98, Bristol, UK, 1998, pp. 49–54
15 ITU-T Recommendation P.910: 'Subjective video quality assessment methods for multimedia applications'. International Telecommunication Union, Geneva, Switzerland, 1999
16 Barnett, V.: 'Sample survey principles and methods' (Edward Arnold, London, 1991)
17 ITU-R Recommendation BT.500-11: 'Methodology for the subjective assessment of the quality of television pictures'. International Telecommunication Union, Geneva, Switzerland, 2002
18 NTIA Report 02-392: 'Video quality measurement techniques'. National Telecommunications and Information Administration, USA, 2002
19 Pinson, M.H., and Wolf, S.: 'A new standardized method for objectively measuring video quality', *IEEE Trans. Broadcast.*, 2004, **50**, (3), pp. 312–322
20 Winkler, S., and Campos, R.: 'Video quality evaluation for Internet streaming applications'. Proc. SPIE/IS&T Human Vision and Electronic Imaging, Santa Clara, CA, 2003 5007, pp. 104–115
21 Winkler, S., and Dufaux, F.: 'Video quality evaluation for mobile applications'. Proc. SPIE/IS&T Visual Communication and Image Processing, Lugano, Switzerland, 2003, 5150, pp. 593–603
22 Vetro, A., Sun, H., and Wang, Y.: 'Object-based transcoding for adaptable video content delivery', *IEEE Trans. Circuits Syst. Video Technol.*, 2001, **11**, (3), pp. 387–401
23 Sundaram, H., Xie, L., and Chang, S.-F.: 'A utility framework for the automatic generation of audio-visual skims'. Proc. ACM Multimedia, Juan Les Pins, France, 2002
24 Walpole, J., Krasic, C., Liu, L., Maier, D., Pu, C., McNamee, D., and Steere, D.: 'Quality of Service Semantics for Multimedia Database Systems'. Proc. IFIP TC2/WG2.6 Eighth Working Conference on Database Semantics - Semantic Issues in Multimedia Systems, New Zealand, 1999, pp. 393–412
25 XingMPEG® encoder: http://www.xingtech.com/
26 Rimell, A., and Hollier, M.: 'The Significance of Cross-Modal Interaction in Audio-Visual Quality Perception'. Proc. IEEE Workshop on Multimedia Signal Processing, Copenhagen, 1999, pp. 509–514
27 Elting, C., Zwickel, J., and Malaka, R.: 'Device-dependent modality selection for user-interfaces - an empirical study'. Proc. Int. Conf. on Intelligent User-Interfaces, 2001, pp. 55–62
28 Snoek, C.G.M., and Worring, M.: 'A review on multimodal video indexing'. Proc. IEEE Int. Conf. on Multimedia & Expo (ICME), Lausanne, Switzerland, 2002, Vol. 2, pp. 21–24
29 Kim, M.B., Nam, J., Baek, W., Son, J., and Hong, J.: 'The adaptation of 3D stereoscopic video in MPEG-21 DIA', *Signal Process., Image Commun.*, 2003, **18**, (8), pp. 685–697
30 Asadi, M.K., and Dufourd, J.-C.: 'Multimedia adaptation by transmoding in MPEG-21'. Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services, Lisboa, 2004
31 Serif, T., Gulliver, S.R., and Ghinea, G.: 'Infotainment across access devices: the perceptual impact of multimedia QoS'. Proc. ACM Symp. on Applied Computing, 2004, pp. 1580–1585
32 Wikstrand, G., and Eriksson, S.: 'Football animations for mobile phones'. Proc. NordiCHI, 2002, pp. 255–258