Intelligent Media Format Conversion for Universal Multimedia Service over Heterogeneous Environments

Kwang-eun Won and Kwang-deok Seo
Yonsei University
Wonju, Gangwon, South Korea
e-mail: kdseo@yonsei.ac.kr

Jae-wook Lee
LG Electronics
Gasandong, Gurogu, Seoul, South Korea
e-mail: suance88@hanmail.net

***Abstract*—The rapidly growing Internet has come with an increasing heterogeneity and diversity in the network devices and connections used for information access. To guarantee the Quality of Service (QoS) to multimedia applications over heterogeneous networks and terminals, content adaptation is one of the most important technologies. The content adaptation can generally be classified into two different categories: 1) content scaling which controls the quality of the content according to the given constraints, 2) media format conversion which converts the given media format to a different one, such as video-to-image conversion. Although the content scaling approach can be employed to match the quality of the content to a worse condition of terminal and network resources, the resulting quality could be significantly deteriorated. In this case, media format conversion could be more preferable and better solution to guarantee the quality of the media service. In this paper, we specifically focus on converting video to image format. The effectiveness of the proposed media format conversion in terms of human perceptual aspect is verified through extensive simulations.**

Keywords-intelligent media format conversion; intelligent multimedia, universal multimedia service; multimedia access over heterogeneous environments, quality of service.

I. INTRODUCTION

In general, universal multimedia service deals with delivery of multimedia content (images, video, audio, and text) over different network conditions, user and publisher preferences, and capabilities of terminal devices. Recently Universal Multimedia Access (UMA) has become a new trend in multimedia communications [1]. In the UMA, the content adaptation is the important process to cope with various constraints of terminal and network. It can be commonly divided into two different categories as shown in Fig. 1: one is the content scaling technique, which controls the quality of the content according to the given constraints and the other is media format conversion, which converts one media format to another, such as video-to-image conversion.

Conventionally, the content scaling approach has been used to match the quality of the content to the corresponding constraints in terminal and network resources. However, media format conversion currently appears to take an important role in the evolution of UMA. In terms of human perceptual information, the quality of the content can be significantly destroyed although the content scaling is able to sufficiently reduce the rate. In this case, instead of the content scaling approach, media format conversion can be more preferable choice to guarantee the Quality of Service (QoS). MPEG-21 [2] provides the conversion preference, so that users can be able to customize the media format to the adapted resource. Besides, MPEG-21 also provides the AdaptationQoS descriptors to enable the adaptation engine to automatically scale the resource to cope with constraints of terminal/network.

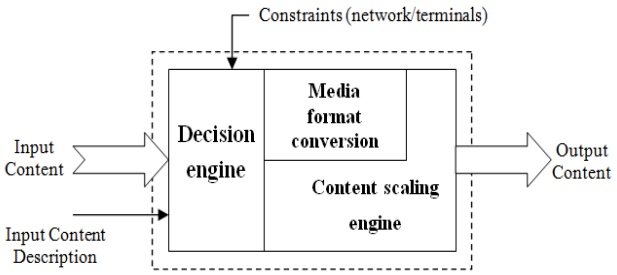


Figure 1. Content adaptation process.

In this paper, we propose an intelligent media format conversion technique, which converts video to image and guarantees the acceptable QoS under the obtained conversion boundary.

This paper is organized as follows. In Section II, we review related works including Overlapped Content Value (OCV) model. Section III describes the proposed intelligent video-to-image format conversion technique. In Section IV, we show the experimental results to verify the effectiveness of the proposed media format conversion in terms of human perceptual aspect. Finally, concluding remarks are presented in Section V.

II. REVIEW OF RELATED WORKS

Content adaptation can be performed either at the client, at an intermediate proxy side, or at the server [1]. Usually, most content adaptation systems [3]-[5] are proxy-based.

 ISBN: 978-1-61208-576-0

Client device requests Web pages. Then, the proxy catches client device's request for Web pages, brings the requested content, and adapts it. Finally, the proxy sends the adapted version to the client.

In the TranSend project [3], a proxy transcodes Web content on the fly. The adaptation which is also called as "distillation," is primarily limited to image compression and reduction of image size and color space. Video is also converted into different frame-rates and encodings using a video gateway.

Bickmore and Schilit [5] also propose a proxy based mechanism. They use a number of heuristics and a planner to perform outlining and elision of the content to fit the Web page on the client's screen. These transcoding proxies typically consider a few client devices and employ static content adaptation strategies. A common policy [3] [5] is to scale all images by a fixed factor. Therefore, these transcoding proxies fail to address the variation in the resource requirements of different Web documents. The set of client devices will also grow more diversity. Certain resources, such as effective network bandwidth, costs and patience of the users can be different for similar client devices. The static adaptation policies used by these systems do not handle well this variability in Web content and client resources.

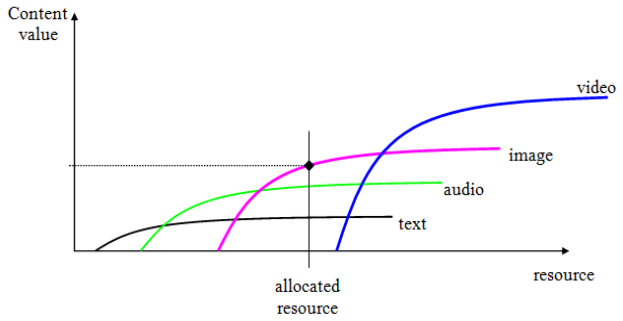


Figure 2. The overlapped content value model of a content item.

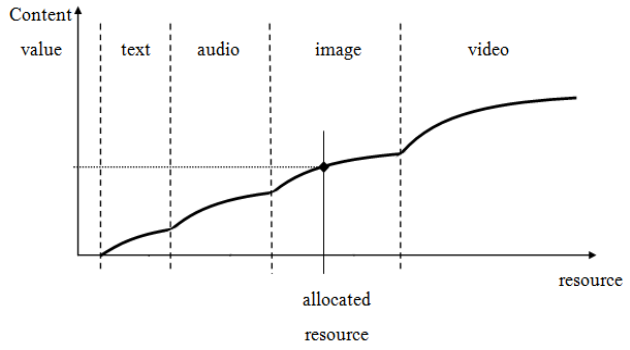


Figure 3. The final content value function of a content item.

Most research results on content adaptation have focused on transcoding of contents within a single format [6] [7] or on a single type of format conversion, e.g., video to images [8]. Media format conversion may be supported in the approach of [9]. However, this approach works with one content item, resulting in little practical use. The approach in [1] is one of few adaptation approaches that can handle multiple formats and multiple contents. However, its resource allocation method is not quite suitable for making decision on media format conversion. Especially, it does not address rapidly changing user/network conditions.

Thang et al. [10] employed the OCV model to represent the content value in different formats according to resource. This model helps finding the conversion boundaries between different formats. The underlying idea is that the conversion boundary between various formats depends on the perceptual qualities of the media formats. Fig. 2 shows the example model of a content that is originally of video format. Here, the content value curve of each format can be assigned manually or automatically. The final content value function of the content is the upper hull of the model, and the intersection points of the curves represent the conversion boundaries between formats. Fig. 3 shows the final content value function and the conversion boundaries of the content. Based on these conversion boundaries, we can quantitatively make the decision on media format conversion so as to maintain an acceptable QoS. The OCV model enables the adaptation engine to determine appropriate media format to be achieved for a given constraint in a quality-aware way. The content value (quality) of a media format can be measured in various utilities. For example, video content value can be computed using both the Peak-Signal-to-Noise-Ratio (PSNR) value and the Mean Opinion Score (MOS) value as follows:

$$Content\_value = z_1 \cdot PSNR\_value + z_2 \cdot MOS\_value \quad (1)$$

where z_1 and z_2 are the appropriate scale factors which are defined by the user, and the sum of them is 1. The criteria for establishing appropriate scaling factors depend on the relative importance between objective quality and subjective quality. If objective quality is comparatively more important than subjective quality, we set z_1 to be greater than z_2. Otherwise, z_1 should be less than or equal to z_2. For more details on computing the OCV model, you are referred to [10]-[12].

III. Proposed Intelligent Media Format Conversion

We propose an intelligent media format conversion technique, which converts the visual information of video content to important image sequence, based on the OCV model in UMA environment. Fig. 4 shows the overall system

 ISBN: 978-1-61208-576-0

architecture of the content adaptation employing media format conversion from low quality video to high quality image. Video-to-image format conversion process is as follows: The decision engine takes the original video content and resource constraints as its inputs. Then, the decision engine analyzes the resource constraint and makes optimal decision between format conversion and video transcoding, so that the adapted video has the most value when presented to the user. The media format conversion engine includes the specific operations to adapt the video according to instructions from the decision engine. In Fig. 4, when the given bit-rate is lower than the conversion point, which is generated by the OCV model, video-to-image format conversion is applied. Otherwise, video is scaled by the video transcoder. Under the situation of very limited bit-rate, video transcoder cannot generate satisfactory video quality. For a given resource constraint (specifically, bit-rate as in Fig. 4), the media format that maximizes the utility corresponding to the convent value should be selected for the universal multimedia service.

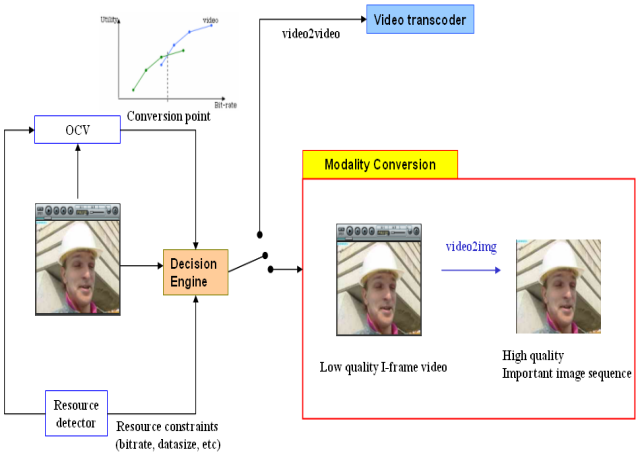


Figure 4. Overall system architecture.

In order to decide the conversion boundary from the video to image, we need to get utility curves: one is video utility curve and the other is image utility curve.

A. *Video utility curve*

At the given bitrate, PSNR value is estimated according to frame dropping and Quantization Parameter (QP) adjustment. Then, we normalize the PSNR value within utility value 1 in order to map *PSNR* value into OCV model. For generating the scaled video, we use the operation which is combined with frame-dropping and requantization. This process results in the video utility curve shown in Fig. 5. The average *PSNR* values and bitrates of the scaled video are measured to provide the video utility curve. In (2), the content value of video (V) is calculated by multiplying the *PSNR* values with the scale factor of w. This scale factor is used to map the video *PSNR* value into the range [0, 1].

$$V = w \times MV_{PSNR}, \quad where\ w = \frac{1}{MV_{PSNR\_\max}}. \qquad (2)$$

The final video utility curve is shown in Fig. 6. From right to left on this curve, each point represents the video streams generated by combining frame-dropping and quantization parameter adjustment. For example, the first point represents the video streams having no frame-dropping with a fixed quantization parameter.

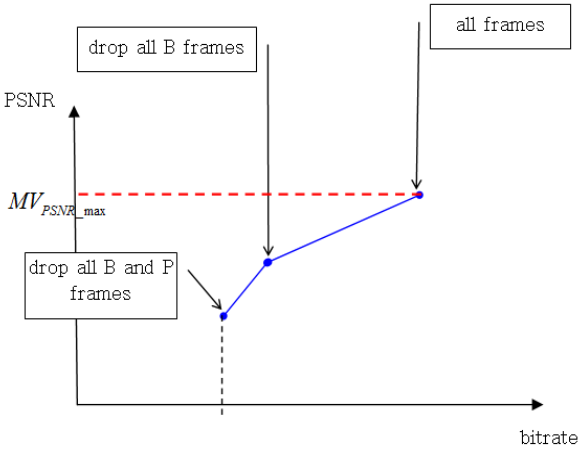


Figure 5. Video utility curve.

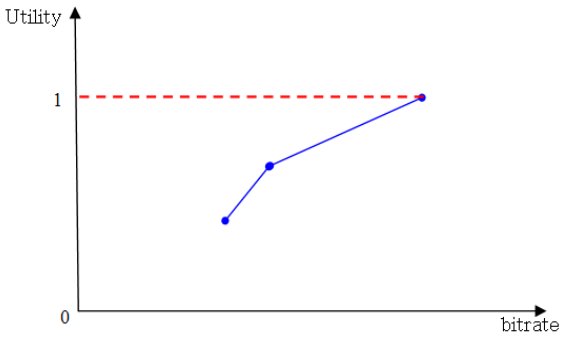


Figure 6. Normalized video utility curve.

B. *Image utility curve*

For convenience, we suppose the data size of important image sequence is approximately equal to an average value. So, the resource amount of image sequence is linearly proportional to the number of important images. Image sequences are extracted from the full sequence of images (decoded from original video) using various methods described in [13] and [14]. An extracted image sequence is said to represent the best possible summary of the video. Thus the scaling operation for image format 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 which ranges from 0 to infinity.

We see that, when the image sequence has all the frames, the maximum content value of image format is the same as the original video, which is 1, and then we can set the scale factor of image format to be 1. The distortion D can be changed into content value as follows.

$$V = 1/(1 + A \cdot D), \quad (3)$$

where A is an unknown constant. It should be noted that (3) is a more generalized case of the formula $V = 1/(1 + D)$ proposed in [1]. The constant A, which actually controls the slope of the image utility 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 [13] as follows:

$$V' = 1/(1 + A \cdot D'). \quad (4)$$

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

$$V' = w \cdot MV'_{PSNR}. \quad (5)$$

From (4) and (5), we have

$$A = \frac{w \cdot MV'_{PSNR} \cdot D'}{1 - w \cdot MV'_{PSNR}}. \quad (6)$$

C. *Mapping of video and image utility curves into OCV model*

Fig. 7 shows the video and image utility curves mapped into OCV model [10]. From this utility curve, we can find the conversion point and the resource constraint where the current format is converted.

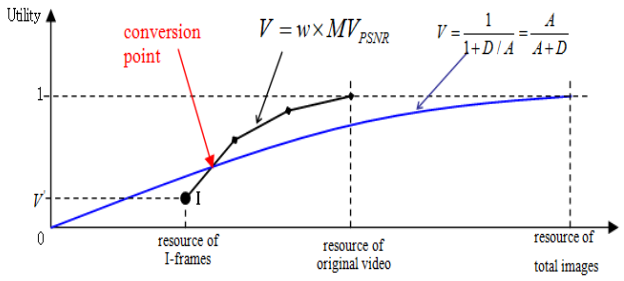


Figure 7. Video and image utility curves mapped into OCV model.

In order to decide the number of extracted images, we propose the decision procedure on the number of extracted images as shown in Fig. 8.

Once we get the decoded JPEG images from input video, we calculate the average of the JPEG image size (Q), which is almost similar to that of an I-frame size. Next, we decide the number of extracted image (N) according to resource constraint (R) and extract the images [13]. Then, we calculate the total data size of the extracted images (T) and the difference value (D) of the total data size of extracted images and the resource constraint (R). In addition, we should check whether the difference value (D) is bigger than zero or not. Then, we need to *update* the number of image to be selected.

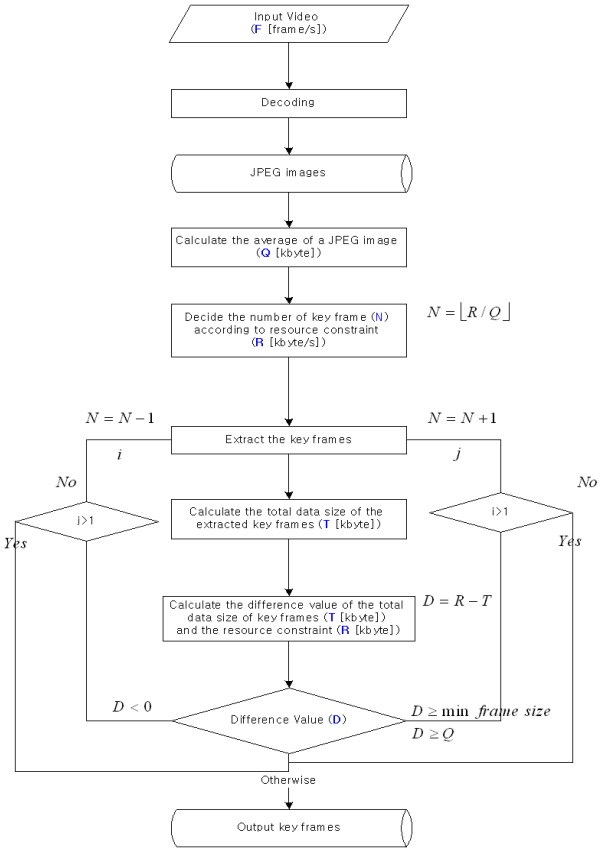


Figure 8. Flow chart for deciding the number of image.

If the difference value (D) is smaller than zero, we decrease the number of image by 1. Otherwise, we increase the number of image by 1. After checking the difference value (D) again, we can find the number of extracted image at last.

IV. EXPERIMENTAL RESULTS

The experiment was performed using a desktop computer system with a Quad-Core 4.0 GHz CPU and 4G Byte

 ISBN: 978-1-61208-576-0

memory. For input video, we used a test video sequence with 720p HD (1280x720) resolution, *English.mpg*, for which H.264 compression is applied with the frame rate of 30 fps. Each GOP consists of 15 frames with the structure IBBPBBPBBPBBPBB. The operations of content scaling and media format conversion in the experiment are carried out off-line. That is, a number of content versions of video and image are stored in advance. The content value of the original video is supposed to be 1, and content value of image is mapped into the range [0, 1]. The final video utility curve is shown in Fig. 9.

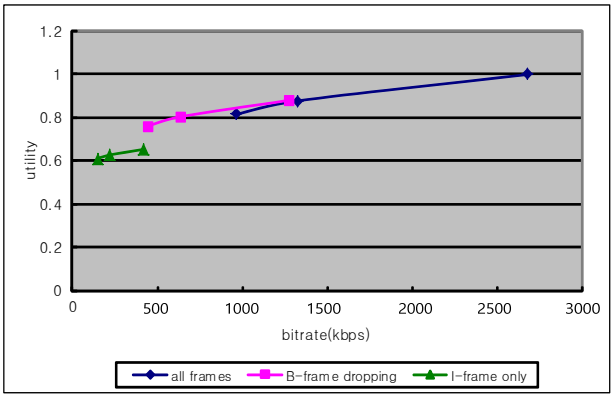


Figure 9. Video utility curve for the test sequence.

When the number of images is 0, the distortion is obviously infinity, and the full image sequence has zero distortion. Also when the image sequence has all frames as the original video, the maximum content value of image format is the same as the original video, which is 1. In order to generate the image utility value, firstly we need to calculate constant A by using (3), (4), and (5). The scale factor w is $1/\max PSNR\ vlue = 1/32.11 = 0.031$ and the distortion of I-frame stream $D^{'}$ is approximately 55.57, which is calculated by the extraction method [13]. Then, its *PSNR* value $MV^{'}_{PSNR}$ is 21.01. Now using (2), we can get constant value A as $A = \frac{w \cdot MV^{'}_{PSNR} \cdot D^{'}}{1 - w \cdot MV^{'}_{PSNR}} = 105.18$.

Using the calculated constant A and distortion value, we are able to draw the utility curve shown in Fig. 10. Also, we can map the video utility curve and image utility curve into OCV model shown in Fig. 11. After mapping two utility curves into OCV model, conversion point *B* can be found at 324.2 kbps as shown in Fig. 11. For the perceptual comparison of each content version, we select point *A* (about 147.8 Kbps) as a comparison point. At that point, the video curve indicates I-frame stream with QP=30, and then it corresponds to 5 important images of the image curve.

Fig. 12 shows two kinds of content versions of video and image at 324.2 Kbps. The video version consists of 20 I-frames with QP=20. Meanwhile, the image version consists of only 8 images with original spatial quality.

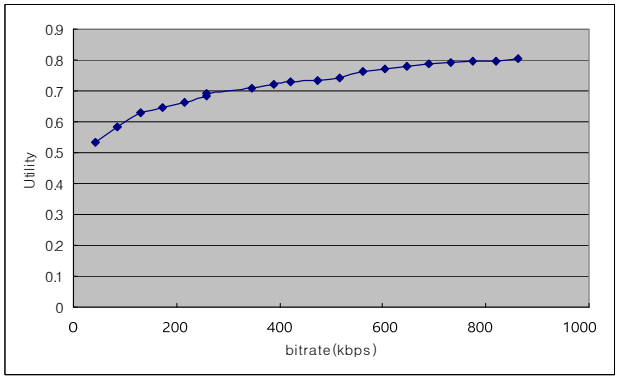


Figure 10. Image utility curve for the test sequence.

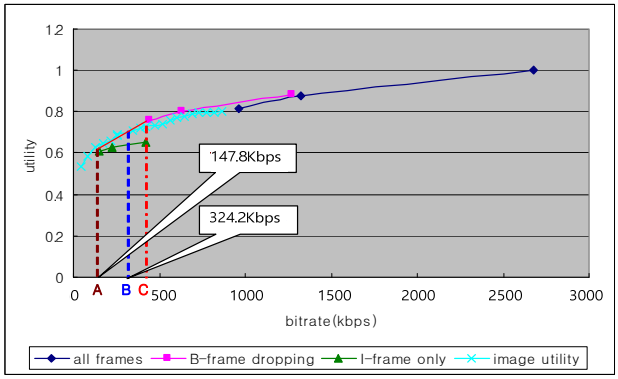
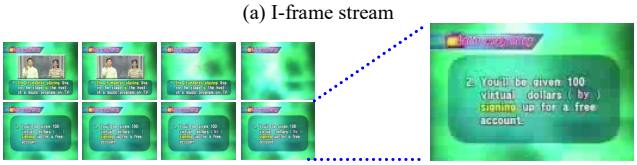


Figure 11. Overlapped Content Value model for the test sequence.

(a) I-frame stream



(b) Image sequence consisting of 8 high quality images

Figure 12. Comparison of the visual content between video and image formats.

In the case of this test video sequence for English education, text is the most important information source. However, we can obtain little information from the

degraded I-frame stream as shown in Fig. 12(a). Therefore, although the image version has fewer high quality images due to spatio-temporal trade-off, we can say that the image version has much better human perceptual information than the video version. From the results, we can see that in terms of human perceptual information, the quality of the video version can be destroyed significantly. So, it is reasonable to select the image version to transmit to the end users at this bitrate.

V. CONCLUSION AND FUTURE WORK

In this paper, we determined the conversion boundary among different media formats, and present media format conversion technique from video-to-image format. This approach guarantees the acceptable QoS under the obtained conversion boundary. Although conventional content scaling method can sufficiently decrease the bit-rate, it could seriously give rise to the destruction of the important visual information, which the original video originally has. In this case, the proposed media format conversion method can be a good alternative solution to avoid this kind of problem. The experimental results demonstrate: 1) it is better to transmit the important image sequence than to send only I-frame video stream since the important image sequence has most of key frames which I-frame video may missing, 2) the scaled video has little perceptual information at the low bitrate below the conversion point because it almost loses text information, which could be the most important information source, due to serious quality degradation. Meanwhile, the selected important image sequence has much valuable human perceptual information even though it includes a small number of high quality images because of trade-off between spatial and temporal quality.

ACKNOWLEDGMENT

This work was supported by a grant 'Biotechnology & GMP Training Project' from the Korea Institute for Advancement of Technology (KIAT), funded by the Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea (N0000961).

REFERENCES

[1] R. Mohan, J. Smith, and C. Li, "Adapting Multimedia Internet Content for Universal Access," *IEEE Trans. Multimedia*, vol. 11, no. 1, pp. 104-114, Mar. 2009.

[2] MPEG MDS Group, "Study of ISO/IEC 21000-7 FCD - Part 7: Digital Item Adaptation," ISO/IEC JTC1/SC29/WG11 N5933, Brisbane, Australia, Oct. 2003.

[3] A. Fox, S. Gribble, Y. Chawathe, and E. Brewer, "Adapting to network and client variation using active proxies: Lessons and perspectives," *IEEE Personal Commun*., vol. 55, 2013, pp. 10-19.

[4] J. Smith, R. Mohan, and C. Li, "Transcoding Internet content for heterogeneous client devices," in Proc. *Int. Symp. Circuits and Systems (ISCAS)*, Monterey, CA, 1998, pp. 1-10.

[5] T. Bickmore and B. Schilit, "Digestor: Device-independent access to the World Wide Web," in Proc. *Int. WWW Conf*., Santa Clara, CA, 2007, pp. 27-35.

[6] N. Bjork and C. Christopoulos, "Video Transcoding for Universal multimedia Access," in Proc. *ACM Multimedia*, pp. 75-79, Nov. 2000.

[7] K. Lee, H. S. Chang, S. S. Chun, H. Choi, and S. Sull, "Perception-based image transcoding for universal multimedia access," *Int. Conf. on Image Processing*, pp. 475-478, 2011.

[8] T. Kaup, S. Treetasanatavorn, U. Rauschenbach, and J. Heuer, "Video analysis for universal multimedia messaging," *IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 211-215, 2002.

[9] W. Lum and F. Lau, "A QoS-sensitive content adaptation system for mobile computing," *Computer Software and Applications Conference*, pp. 680–685, 2016.

[10] T. Thang, Y. Jung, Y. Ro, J. Nam, M. Kimiaei, and J. Dufourd, "CE report on Modality conversion preference-Part I," ISO/IEC JTC1/SC29/WG11 M9495, Pattaya, Thailand, Mar. 2003

[11] S. Chandra and C. Ellis, "JPEG compression metric as a quality aware image transcoding," in Proc. *USENIX Symp. Internet Technologies and Systems*, Boulder, Colorado, Oct. 2009.

[12] J. Kim, Y. Wang, and S. Chang, "Content-adaptive utility based video adaptation," in Proc. *Int. Conf. on Multimedia & Expo (ICME)*, 2013.

[13] H. Lee and S. Kim, "Iterative Key Frame Selection in the Rate-Constraint Environment," *Image Communication*, Issue 28, pp. 1-15, 2013.

[14] H. Chang, S. Sull, and S. Lee, "Efficient Video Indexing Scheme for Content-Based Retrieval," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 19, pp. 1269-1279, Dec. 2009.

 ISBN: 978-1-61208-576-0