MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Towards Maximizing the End-User Experience

Ajay Divakaran, Anthony Vetro, Takashi Kan

TR2004-055 June 2004

Abstract

Increasing bandwidth and diverse devices have led to a proliferation of options for the consumer of digital video. The user can easily be overwhelmed by the diversity of devices, bandwidth and content. In our view, maximizing the end-user experience involves first helping the user to easily search for and retrieve desired content, and then delivering the content to him by seamlessly adapting to his device and available bandwidth. The first part is therefore based on semantic criteria while the second part is baased on purely signal-based criteria. We propose that scalable content summarization combined with dynamic content adaptation, and unobtrusive processing of user preferences comprise the key technologies for maximizing the end-user experience.

ICME 2004

Proc IEEE Int’l Conference on Multimedia and Expo (ICME)

Towards Maximizing the End-User Experience

Ajay Divakaran, Anthony Vetro and Takashi Kan
Mitsubishi Electric Research Laboratories
201 Broadway, Cambridge, MA 02139, USA
{ajayd,avetro,kan}@merl.com

Abstract

Increasing bandwidth and diverse devices have led to a proliferation of options for the consumer of digital video. The user can easily be overwhelmed by the diversity of devices, bandwidth and content. In our view, maximizing the end-user experience involves first helping the user to easily search for and retrieve desired content, and then delivering the content to him by seamlessly adapting to his device and available bandwidth. The first part is therefore based on semantic criteria while the second part is based on purely signal-based criteria. We propose that scalable content summarization combined with dynamic content adaptation, and unobtrusive processing of user preferences comprise the key technologies for maximizing the end-user experience.

1. Introduction

Consumption of digital media is rapidly changing from the typical "TV in the living room" to ubiquitous access. Advances in connectivity have enabled reception and/or consumption of digital media on diverse devices such as mobile phones, PDA's, and portable players. Furthermore, since the consumers are mobile, the available bandwidth and network characteristics are constantly changing depending on the location of the user. The challenge posed by ubiquitous access is therefore to satisfy user needs while making the diversity of devices and networks completely transparent. In other words, the user should have a wide range of options, and yet be given a seamless experience despite the constantly varying device-network characteristics.

In our past work [1], we have proposed a combination of summarization and transcoding to serve ubiquitous multimedia access. This paper proposes enhancements over this previous work. A key technical advancement of our proposed summarization is that it yields scalable summaries that allow simple and quick adjustment of summary length as per user requirements. This paper also considers the role that such standards as MPEG-7 and MPEG-21 play towards fulfilling the stated goals. In [2], Tseng, et al. describe how these standards may be used for personalizing video. In this paper, we focus on specific techniques towards maximizing the user experience.

2. User Needs & Constraints

The essential objective of ubiquitous multimedia access is to give the user what he wants, when he wants and where he wants it. We term giving the user what he wants as a *semantic need* that is independent of the terminal-network characteristics. However, providing the content when and where he wants highly depends on the terminal-network characteristics. We term this an *infrastructure constraint*. We propose that we can tackle these needs separately and provide a satisfactory solution to the overall problem.

Figure 1 provides a high level abstraction of the factors and operations to achieve a maximum user experience. As shown in the subfigure on the right, the operations generally include processing, adaptation and presentation, while factors that affect these operations include the semantic needs, infrastructure constraints and the multimedia itself including the content and possibly its corresponding metadata. If all needs and constraints are properly accounted for, a maximum user experience could be achieved. In the subfigure on the left, an example representation considering the dimensions of space, time and granularity of content is shown.

An important point to make is that the notion of a user experience is not well defined at present. In [3], it is suggested that human factors will play a significant role in maximizing the quality of

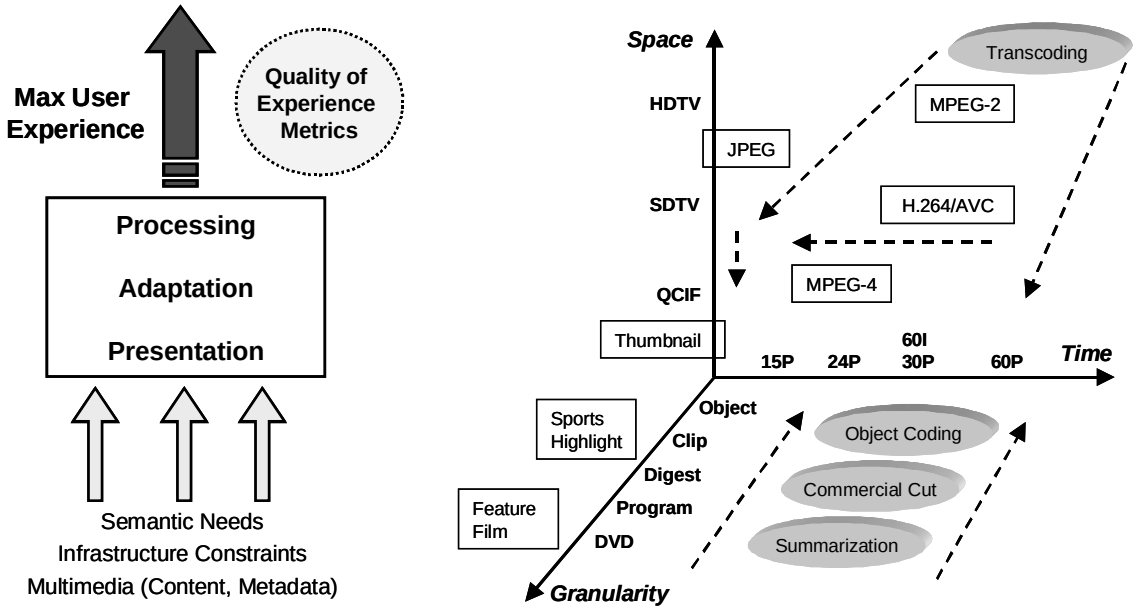

Fig. 1. Factors and operations to achieve a maximum user experience. [Left] General illustration of concept. [Right] Example representation with dimensions of space, time and granularity of content, where unshaded boxes indicate particular formats, dashed lines indicate traversal of content formats within space, and shaded ovals indicate techniques used to achieve output.

experience. While we strongly agree with this, much of the technology to realize this vision is not fully mature. Furthermore, there is a lack of metrics to define an optimal user experience. This paper takes a more pragmatic approach that emphasizes several important factors that should be considered in maximizing the user experience. We believe that these factors will remain significant even after human factors are better understood and exploited.

3. Satisfying Semantic Needs

Searching and browsing video has been an active area of research and is one of the primary applications of the MPEG-7 standard [4]. Retrieval from still image databases, based on higher and/or lower level features, relies largely on indexing, automatic or otherwise. We propose that consumer video retrieval cannot rely on indexing alone. Since video has a temporal dimension, it would take too much time for the consumer to go through the results of a query. Note that the consumer is primarily interested in retrieving video programs. These are usually at least 30 minutes long, which implies that it would take 30**N* minutes to go through *N* results of a query. Therefore, there is a clear need for video summarization to enable the consumer to go through digests or abstracts of the program before deciding to watch it in its entirety. Furthermore, the consumer could also use video summaries as entry points into the content itself if such interactivity is feasible, such as in a home video server.

Note that the mobile consumer in particular is not likely to watch video over extended periods of time. He is much more likely to consume snippets since watching an entire feature film on a matchbox sized display might not be feasible or enjoyable. Therefore, it is imperative for the consumer to have the ability to retrieve desired video and then play a highly abbreviated version thereof.

We have found that content-based video summarization works best when the semantic context has already been narrowed. For example, summarizing 1000 hours of video through purely content-based methodology would still result in more video than a user can handle. However, typically there is a lot of available textual meta-data at the program level which is typically at a granularity of half an hour or so. Our approach is therefore to combine keyword-based indexing at the program level with content-based summarization of the video retrieved by the keyword-based indexing. In this way, we first search at the program level through textual, i.e., conceptual, search and then once the retrieved video has been narrowed to a reasonable length, such as say 10 hours, we can then use content-based video summarization to skim the video.

To provide the consumer with maximum flexibility, the video summarization has to be flexible as well. From a content abstraction point of view,

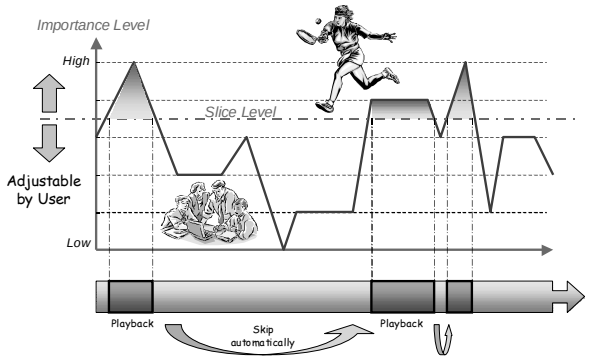

Fig 2. Scalable summarization using importance levels.

flexibility translates to the ability to obtain a video summary of any desired length. In other words, the summaries should be scalable.

We propose obtaining scalable summaries by assigning an importance level to each portion of the content. Then, as illustrated in Fig. 2, only the portions of the content with importance greater than a certain threshold can constitute the summary. The desired length is therefore achieved by choosing an appropriate importance level threshold.

In past work [5], we have used the length of the crowd reaction (applause and/or cheering) in soccer game, to assign importance levels to portions of soccer, golf and baseball games, and found that we could vary the length of the summary successfully. In ongoing work, we have obtained promising results on the automatic extraction of such importance levels for various genres of content using key-event and key-frame extraction. Note that there need not be any real-time constraint on the extraction of the importance levels even though quick computation is of course a bonus. Therefore, semi-automatic importance level generation by professional content creators is a feasible and effective option.

Finally, note that the summary does not necessarily have to consist of video snippets. However, displaying key-frames/thumbnails or other variations of the video is relatively straightforward, so we do not cover those options in this paper.

4. Satisfying Infrastructure Constraints

Once the content has been located and sufficiently narrowed as described in the previous section, i.e., through a combination of indexing and summarization, the next challenge is to successfully play the content on the consumer's device regardless of its location. In other words, the content needs to be adapted to the usage environment, in particular the terminal capabilities and network characteristics. Since the conditions are likely to change over time, the transcoding techniques we employ are dynamically adjustable so as to give the user a seamless experience. In the following, these two aspects of the adaptation are discussed, namely the usage environment descriptions, which are standardized in Part 7 of the MPEG-21 standard, Digital Item Adaptation (DIA) [6], and the dynamic transcoding operation.

4.1 MPEG-21 Usage Environment Descriptions

The usage environment description tools as specified in MPEG-21 DIA include the description of terminal capabilities and network characteristics. In the following, select components of these description tools are discussed. Further details on other usage environment descriptions may be found in [7].

Two important terminal capabilities are decoding and display capabilities. The decoding capabilities specify the format a particular terminal is capable of rendering, e.g., MPEG-4 Simple Profile at Level 3. Under display capabilities, attributes such as the size, color depth and resolution of the display are specified. Given the variety of different content representation formats that are available today, it is necessary to be aware of the formats that a terminal is capable of decoding. Furthermore, to ensure that images/video can be properly rendered on the display of the device, the size, resolution and color depth are particularly important.

Two main categories are considered in the description of network characteristics, network capabilities and network conditions. Network capabilities define static attributes of a particular network link, while the conditions describe more dynamic behavior and include timing information as well. As part of the network capabilities are the maximum capacity of a network and the minimum guaranteed bandwidth that a particular network can provide. Network conditions specify error, delay and available bandwidth. The error is specified in terms of packet loss rate and bit error rate. Several types of delay are considered including, one-way and two-way packet delay, as well as delay variation. Available bandwidth includes attributes to describe the minimum, maximum, and average bandwidth of a particular link. One application of these descriptions is to improve the transmission efficiency of a video stream by increasing the rate of intra-coded blocks if the packet loss rate is high.

4.2 Dynamic Transcoding

A smooth transition of quality due to a changing network bandwidth is one factor to ensure the best user experience. Sudden changes in picture quality or the excessive dropping of frames could lead to disturbing artifacts in the perceived video. While a number of metrics have been proposed to gauge the picture quality, e.g., in terms of blocking artifacts, there is still no suitable measure that accounts for perceived temporal distortion due to frame dropping. Of particular importance is when frames are dropped in a non-uniform manner, which may lead to a jerky rendering of the video. The exploration of such measures is currently an active area of research.

With the above considerations in mind, we propose a rate control scheme in the transcoder that continually accounts for changes in network conditions periodically, e.g., at the GOP boundary of the input stream. Regardless of whether the input stream has a GOP structure or not, rate allocation is typically performed over a period of frames. Our system has the ability to change the output bit-rate as well as the output frame-rate at the boundary of each period. We do not drop frames dynamically to avoid jerkiness in the video, but rather increase or decrease the output frame-rate within each period. Based on this decision, the allocation to output frames within that period can be made. Constraints are imposed on both the bit-rate and frame-rate so that a smooth transition could be achieved.

4.3 Optimizing Multiple Constraints

Making optimal trade-offs in spatial and temporal quality is still an open area of research. However, tools have been standardized by MPEG-21 DIA to assist with this objective. In particular, the AdaptationQoS tool specifies the relationship between the constraints, feasible adaptation operations satisfying the constraints, and associated utilities (qualities). The utility values may be in the form of either subjective or objective measures and may be derived based on different adaptation operations and the parameters of those operations. A wide variety of possible constraints, adaptation operations and utility metrics are defined by the standard to enable a rather rich set of relationships to be expressed. At the moment, we consider terminal capabilities and network characteristics as the primary constraints, but in the future it is feasible to explore a variety of other factors such as the constraints on the consumer's time, power being consumed at the receiver, computational load in the server or proxy, and so on. To satisfy such constraints as consumer time, dynamic transcoding and scalable summarization techniques as described in this paper could be used.

5. Concluding Remarks

In the proposed approach to maximizing the end-user experience, the consumer only deals with the content search interface and does not have to explicitly deal with the terminal capabilities and network characteristics. Such a division of responsibility is reasonable since the consumer should make the determination whether the content is desired or not. The content adaptation on the other hand has many lower level details that would overwhelm the consumer. We thus ensure that the end-user experience is maximized by (a) helping the user conveniently locate and browse desired content, (b) making the adaptation to usage environmental changes completely transparent, and (c) ensuring smooth transitions in quality due to changes in network conditions. While there are certainly other factors that should be accounted for in measuring the quality of experience, we believe that the above factors play a significant role and can be practically accounted for today. We plan to present a demonstration and experimental results at the conference.

References

[1] A. Vetro, A. Divakaran and H. Sun, Providing Multimedia Services to a Diverse Set of Consumer Devices, *Proc. IEEE Int'l Conf. Consumer Electronics*, Los Angeles, CA, June 2001.

[2] B.L. Tseng, C.Y. Lin and J.R. Smith, "Using MPEG-7 and MPEG-21 for Personalizing Video," *IEEE Multimedia*, January-March 2004.

[3] F. Pereira and I. Burnett, "Universal Multimedia Experiences for Tomorrow," *IEEE Signal Processing Magazine*, vol. 20, pp. 63-73, March 2003.

[4] ISO/IEC 15938, Information technology – Multimedia Content Description Interface.

[5] A. Divakaran, K. A. Peker, R. Radhakrishnan, Z. Xiong and R. Cabasson, "Video Summarization using MPEG-7 Motion Activity and Audio Descriptors," Video Mining, Eds. A. Rosenfeld, D. Doermann and D. DeMenthon, Kluwer Academic Publishers, 2003.

[6] ISO/IEC 21000-7 FDIS, "Information Technology – Multimedia Framework – Part 7: Digital Item Adaptation," Eds. A. Vetro, C. Timmerer and S. Devillers, December 2003.

[7] A. Vetro, "MPEG-21 Digital Item Adaptation: Enabling Universal Multimedia Access," *IEEE Multimedia*, January-March 2004.