Subjective Preference of Spatio-Temporal Rate in Video Adaptation Using Multi-Dimensional Scalable Coding

Yong Wang[*], Shih-Fu Chang[*], Alexander C. Loui[#]

[*] *Department of Electrical Engineering, Columbia University, New York, NY, 10027*
[#] *Imaging Science and Technology Lab, Eastman Kodak Company, Rochester, NY, 14650*
[*]*{ywang, sfchang}@ee.columbia.edu,* [#]*alexander.loui@kodak.com*

Abstract

Video adaptation allows for direct manipulation of existing encoded video streams to meet new resource constraints without having to encode the video from scratch. Multi-dimensional scalable coding such as motion-compensated subband coding (MCSBC) offers an effective and flexible representation for video adaptation. In order to develop robust criteria for selecting optimal spatio-temporal rates used in adaptation, knowledge about subjective preference of spatio-temporal rates is needed. In this paper we study the optimal temporal frame rate over a wide range of bandwidth (50 Kbps to 1Mbps) using subjective quality evaluation with 128 clips and 31 subjects. We analyze the results using statistical testing methods and investigate the dependence of optimal frame rate on user, bandwidth, and video content characteristics. Our findings indicate the agreement among most users and existence of switching bandwidths at which preferred frame rates change. Dependence of the preference on video content types is also revealed. [1]

1. Introduction

Video adaptation is important for universal media access (UMA) applications in which various access environments and platforms impose diverse resource constraints. Video adaptation allows for direct manipulation of existing encoded streams without having to re-encode the video from scratch. Recently, multi-dimensional coding like [1] has shown promises with superior quality compared to conventional DCT-based coding. It also offers great flexibility for video adaptation in multiple dimensions, which is important for UMA because it can provide more flexibility in reshaping media content to achieve better quality delivery compared with single dimensional one. This is true especially for multi-dimensional resource constraint cases (bandwidth, power assumption, computing capability, image resolution, etc.) such as the applications in the handheld devices.

Specifically, spatial adaptation is achieved by recoding the quality of each video frame, while temporal adaptation is used to lower the temporal frame rate. The combination of both can be used to meet a wide range of resource constraints and quality requirements. Despite the availability of such emerging tools, two fundamental questions remain to be answered – (1) how to evaluate the subjective video quality after spatio-temporal adaptation and (2) how to predict the optimal adaptation operation achieving the highest subjective quality.

The most frequently employed video quality evaluation metric is PSNR, which is based on the MSE calculation. In [2], we reported a content-based prediction system to automatically select the optimal frame rate for MC-DCT based video based on the PSNR quality metric. In the literature, there are also analytical models for video quality estimation [3,4]. However, neither PSNR nor the existing models are suitable for a spatio-temporal adaptation scenario because they are not designed to evaluate the subjective quality of videos obtained by using different temporal adaptation operations.

In selecting the optimal spatio-temporal operation, recent work in [5] ran subjective experiments to find the frame rate preferences in low bit rate video coding. They concluded consistent preference of 15fps for low bit rate cases, and offered an explanation based on the motion behaviors of the video content. However, other video content characteristics like spatial complexity were not considered. Such correlation with spatial attributes has been observed in [9] using the adaptation experiments over MDSBC videos.

Another objective of our study is to obtain a quantitative model enabling automatic selection of optimal adaptation options at any given bandwidth for any video. In our prior work [6] an empirical rule about the optimal adaptation frame rate was observed based on MPEG-4 Fine Granularity Scalable coded videos. The rule indicates that human subjects prefer more spatial details when the PSNR quality is below some threshold. Once the threshold of spatial details is met, videos with smoother motion perception, i.e., with a higher frame rate, are preferred. [6] stops short in answering the question about the quantitative boundaries (in terms of bandwidth) beyond which temporal details need to be enhanced.

To address the above challenging issues, we conduct a subjective experiment that evaluates the subjective quality of 128 video clips over diverse bandwidths (6 different bandwidths from 50 Kbps to 1 Mbps). We apply formal statistical testing methods to analyze the dependence of

[1] This work has been supported by Imaging Science and Technology Lab, Eastman Kodak Company.

spatio-temporal preferences on users, video content characteristics and bandwidth. The findings are very informative, indicating the existence of consistent switching bandwidths, about 440Kbps and 175Kbps, at which preferred temporal rates change. In addition, such switching bandwidths also strongly depend on the type of video content characteristics like motion, spatial complexity, etc. In a separate paper [9], we have developed a content-based system with high accuracy in predicting the optimal frame rate for any video.

The rest of this paper is organized as follows. In Section 2, adaptation options in spatial and temporal dimensions are discussed. The subjective experiment setup is described in Section 3. Analysis of the experiment results is presented in Section 4. Section 5 concludes the paper.

2. Codec and Spatio-Temporal Adaptation

In this paper we adopt the motion compensated wavelet/subband video coding system (MCSBC) as our codec platform. Although the codec choice will affect the numerical results, the evaluation and analysis methodology is general and can be extended to other types of codec. MCSBC is an active research topic because of its flexibility for providing multi-dimensional adaptation operations and superior coding quality compared with traditional DCT-based codec, such as MPEG and H.26x [1]. In MCSBC, the video signal undergoes octave subband decomposition in both spatial and temporal dimensions. The coefficients are organized in a 3-dimensional bitplane-based bit stream. The spatio-temporal adaptation is achieved by truncating bitplanes from least significant bits, throwing away high frequency temporal layers, or a combination of both. We do not consider the spatial resolution scalability since the quality degradation in this dimension has been shown to be larger than the others.

In order to meet the bandwidth constraint, in practice the temporal rate is determined in advance, and the spatial adaptation is subsequently run to satisfy the target bit rate. Accordingly, given a target bit rate an adaptation method can be uniquely defined by specifying the temporal adaptation operation t that keeps certain temporal layers. Although the temporal layer offers a finer granularity in adaptation, we consider only three discrete values for $t:\{t_0,t_1,t_2\}$, corresponding to "no temporal adaptation (Full frame rate)", "one-level adaptation (half frame rate)", and "two-level adaptation (quarter frame rate)" in turn. Note given a target rate, multiple solutions using different temporal rates exist.

3. Experiment Setup

3.1 Video pool construction

The video pool consisted of three parts: standard test sequences such as *Akiyo*, *Foreman*, *Paris* and *Mobile*; test sequences used by Video Quality Experts Group (VQEG) [7]; and clips taken from commercial movies. Totally there were 128 video clips. All clips were 288-frame long[2], with CIF (352×288) resolution and an original frame rate of 30fps. They covered a wide range of content characteristic, providing a suitable set for our study in content variation. Also we ensured content consistency within each clip and no shot boundary existed. All of the clips were coded using the MC-EZBC codec [1] with a GOP size of 16 frames. The bandwidths tested in the experiment were $R=\{50,100,200,400,600,1000\}$ Kbps, covering a wide range of bandwidth, with emphasis on the low bandwidth area. Note for different streams, the lowest achievable bandwidth through adaptation vary.

3.2 Subjective experiment

Subjective evaluation of video quality was carried out in a quiet, separated room. The video clips were displayed on a 19" Dell P991 Trinitron monitor at a resolution of 1280×960. Viewing distance was fixed at 5 times the picture width. Totally 31 subjects participated in the experiment. They were undergraduate and gradate students at Columbia University from different departments. Due to the large volume of the evaluation, the video pool was divided into 8 groups, each with 16 distinct clips. Each video group was assessed by 5 subjects (some subjects were enrolled in more than one content group voluntarily).

We adopted the double stimulus impairment scale (DSIS) recommended by ITU-R standard [7] with minor revision. For each video clip and a specific bandwidth, four display windows were aligned in two rows and two columns. The left-top window was for the reference sequence. The other three windows were for the adapted clips. These three clips were adapted to the same bandwidth with full frame rate (30fps), half frame rate (15fps) and quarter frame rate (7.5fps) using spatio-temporal adaptation defined by the temporal adaptation options t_0,t_1,t_2 respectively. The sequential order of different video-bandwidth combination and the layout of the three adapted clips are randomized to prevent evaluation bias. For each adapted clip, the user compares its quality against that of the reference clip and gives a Degradation Mean Opinion Score (DMOS) based on a scale from 1 to 10, corresponding to quality from "very annoying" to "imperceptible".

[2] Some clips from VQEG could only provide 240 frames (15 GOPs).

3.3 Statistical data analysis

The subjective score is a function of video (v), bandwidth (b), temporal frame rate (t), and user (u), i.e., $s(v,b,t,u)$. There are many interesting questions we can answer using the statistics of the scores. For example, for each pair of (v,b), we have multiple users evaluating the quality of videos with different temporal rates (t_0,t_1,t_2). We apply the standard hypothesis testing techniques such as paired-t-test [8] to estimate the confidence in claiming that users have preference in one temporal rate over the other. If we set a higher threshold for claiming quality differences, there will be more cases we cannot draw distinctive conclusions about the preferences and thus the temporal rates being compared will be said to be "tied". Since the number of subjects available for comparing different temporal options for the same (v,b) combination may be small, usually the obtainable conference score is not very high, about 0.75. Paired t-test is found to be adequate for such cases when the number of samples is not large. A shown in Figure 1, setting the confidence threshold to be 0.75 makes the percentage of ties to be about 25%, which is a reasonable proportion.

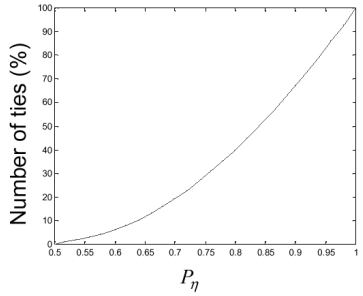

Figure 1: Number of ties v.s. the confidence threshold used in claiming one temporal rate is better than another.

4. Experiment Results and Analysis

We attempt to answer the following questions by analyzing the statistics of the subjective scores.

Q1: Are there different user behaviors in terms of temporal-rate preferences?

In addition to the 128 test clips, we included 3 baseline clips that were seen by all 31 subjects. We hope to use these three clips to assess the consistence of preferences among users. The 3 common clips are of diverse content characteristics. Each clip is tested at 6 different bandwidths. For each video-bandwidth pair, each user assigns subjective scores of different temporal rates – resulting in an 18-dimensional score vector for each user over the baseline video set. The correlation matrix of the score vector for all users was calculated and visualized in Figure 2, in which users were re-sorted based on their mutual similarity. From the figure we can see that most of users (within the dashed region) behaved similarly with high or medium correlation, with others (about 5 users) behaving in a relatively dissimilar way. In other words, this indicates there is a high degree of agreement among preferences by a great majority of users. In the subsequent analysis, we include all the scores from all 31 subjects over the 128 test clips, without attempting to filter out the 5 users as outliers.

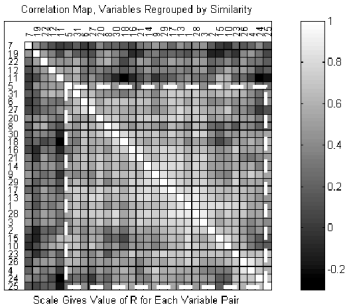

Figure 2: Correlation matrix for assessing user behavior consistence

Q2: Preferred temporal rate at any given bandwidth

Figure 3 shows the number of videos favoring each frame rate at each bandwidth. The determination of temporal rate preference was based on the statistical testing method described in Sec. 3. In the case of ties, the count of each temporal option is increased by a half unit. From the curves, it is clear to see the following trend from high, medium to low bandwidth – the optimal frame rate shifts from full frame rate, half frame rate to quarter frame rate gradually as bandwidth is decreases. Such a trend is intuitive and re-confirms earlier findings. However, the curves also reveal a very important point – there exist two switching bandwidths r_{s_1}, r_{s_2} at about 200Kbps and 450Kbps at which the preferred frame rate changes. If we do not consider video content variation, we can reasonably select the optimal frame rate for adaptation based on these two switching bandwidths. The optimal frame rate for video adapted above r_{s_2} is 30 frames per second. Human subjects prefer half frame rate (15 fps) when the bandwidth is below r_{s_2} but above r_{s_1}. When the bandwidth drops below r_{s_1}, the preferred frame rate becomes quarter rate, i.e., 7.5 fps. Due to the limited data sampling, it is hard to conclude the exact values of the switching bandwidths, but we still can confirm the reasonable range (e.g., r_{s_2} is located in [400, 500] Kbps), which is useful information during adaptation.

Figure 3 reveals the optimal operation behavior in a video adaptation scenario, which is different from the video encoding circumstance as discussed in [5]. In an adaptation scenario, the obtainable quality of the reshaped video stream is restricted by freedom provided by the adaptation techniques, particularly in the low bandwidth area. In Figure 2, for example, at the bandwidth of 100Kbps, it indicates that the operation with frame rate of 7.5fps wins, while in a video encoding case, preference of 15fps is reported in [5]. The difference may be attributed

to the difference in the codecs used. Especially, for video adaptation, the permissible operations are restricted to the coefficient bitplane dropping, temporal layer dropping or combinations of both. For encoding, video bit streams are optimized using all the strategies available for encoding. Despite the above dependence on codecs, we stress the generality of the evaluation methodologies and the analysis strategy.

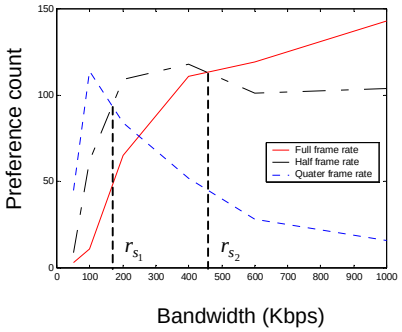

Figure 3: Histogram of the operation preference

Q3: Dependence of optimal temporal rate on video content

We partition all of the videos into categories according to their content complexity – low, medium, and high. The partition was very straightforward: the clips were categorized according to their minimal achievable bandwidth r_{MAB}, defined as the bandwidth below which the adaptation operation can't generate a valid bit stream due to overhead costing coding motion vectors and stream syntax. In our case, r_{MAB} had three distinct values: 50, 100 and 200Kbps. Figure 4 shows the breakdowns of histogram curves shown in Figure 3 into three sets of curves, one for each complexity category. We can clearly see the switching bandwidths shift to the right as the complexity increases. This is quite understandable: more complex videos need more bits for spatial details. Also note that for high-complexity category, there are two comparable winners at both high bandwidth and low bandwidth regions – possibly due to the difficulty in gauging the subjective preferences in a consistent way for this category of videos. In [9], we report a content-based system that can accurately predict the optimal temporal rate at any bandwidth using the content features extracted from the video sequences.

5. Conclusions

In this paper, we study the issue of optimal spatio-temporal tradeoff for video adaptation. We develop a general evaluation method and analysis strategy, but test on videos encoded in specific multi-dimensional scalable format. We apply formal statistical testing method to analyze the trend/consistence of preferences among users, content, and bandwidth. Findings indicate the agreement among users and the existence of two important switching points at about 440Kbps and 175Kbps which define multiple bandwidth regions requiring different optimal frame rate for adaptation.

Reference:

[1] P. Chen and J. W. Woods. "Bidirectional MC-EZBC With Lifting Implementation". IEEE Trans. on Circuits and Systems for Video Technology, 2003. To appear.

[2] Y. Wang, J.-G. Kim, and S.-F. Chang, Content-based utility function prediction for real-time MPEG-4 transcoding, ICIP 2003, September 14-17, 2003, Barcelona, Spain.

[3] VQEG, Final report from the video quality experts group on the validation of objective models of video quality assessment, March 2000. Available at http://www.vqeg.org.

[4] M. A. Masry, S. S. Hemami, "CVQE: A Continuous Video Quality Evaluation Metric for Low Bit Rates," Proc. SPIE Human Vision and Electronic Imaging 2003, San Jose, CA, January 2003.

[5] Yadavalli, G., Masry, M. and Hemami, S.S., "Frame Rate Preferences in Low Bit Rate Video," IEEE Intl. Conf. on Image Processing, Barcelona, Spain, 2003

[6] R. K. Rajendran, M. van der Schaar, S.-F. Chang, "FGS+: Optimizing the Joint Spatio-Temporal Video Quality in MPEG-4 Fine Grained Scalable Coding," IEEE Intl. Symp. on Circuits and Systems, Phoenix, AZ, May 2002.

[7] Methodology for the Subjective Assessment of the Quality of Television Pictures, Recommendation ITU-R BT.500-10, ITU Telecom. Standardization Sector of ITU, August 2000.

[8] Richard Lowry. Concepts and Applications of Inferential Statistics. Online statistic textbook. http://faculty.vassar.edu/lowry/webtext.html

[9] Y. Wang, S.-F. Chang, A. Loui, "Content Based Prediction of Frame Rate Preference in MDSBC Video Adaptation Over a Wide Bandwidth Range," ADVENT Technical Report #202-2004-2 2004.

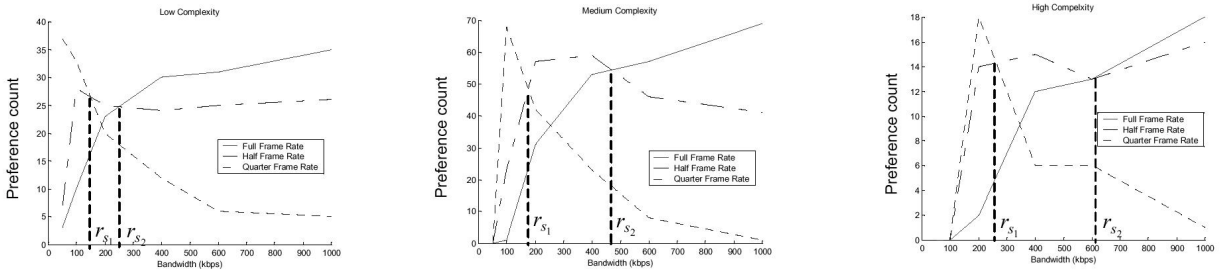

Figure 4: Breakdowns of temporal rate preference distributions into video categories of different complexity