Content-Based Prediction of Optimal Video Adaptation Operations Using Subjective Quality Evaluation

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

[*]*Electrical Engineering Department, 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, which is important in a universal media access scenario. However, currently most of the adaptation research is focused on the issue of predefined video adaptation method and there is little work on systematical solution in selecting optimal operation among multi-dimensional adaptations. In this paper we utilize a content-based prediction framework to select optimal spatio-temporal video adaptation operation using subjective quality evaluation. Content features in the compress domain are employed to explore the correlation between the video content characteristic and optimal adaptation behavior during a pattern classification procedure. The experiment result indicates that our proposed method can efficiently predict the right adaptation operation matching subjective quality evaluation.[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 [Chen03] 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 [Chang02]. Nevertheless, currently most of the adaptation research is focused on the issue of predefined video adaptation method (such as requantization, frame skipping, etc) and there is little work on Multi-Dimensional Adaptation (MDA). One difficulty hampering the development of MDA is the lack of a systematical solution in selecting optimal operation in MDA, which is not as straightforward and analytical as the single-dimensional case. In this paper we are trying to explore this issue by considering the selection of optimal spatio-temporal adaptation through a statistical pattern recognition approach.

Spatio-temporal adaptation method provides us two basic freedoms in reshaping video stream. 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. In order to select the optimal spatio-temporal adaptation, we conduct subjective experiment to evaluate the video quality.

In selecting the optimal spatio-temporal operation, recent work in [Yadavalli03] 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 [Wang04] using the adaptation experiments over MCSBC videos.

In our prior work [Raj02] 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. [Raj02] 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, in [Yong04] we conducted 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 applied formal statistical testing methods to

[1] This work has been supported by Imaging Research and Development, Eastman Kodak Company.

analyze the dependence of spatio-temporal preferences on users, video content characteristics and bandwidth and discovered 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 spatio-temporal complexity, etc. In this paper, we efficiently utilize a content-based framework to select the optimal video adaptation operations with high accuracy. The basic idea of this approach is to consider this optimum prediction problem as a typical classification issue and work it out through a pattern recognition procedure, where the low level video content features act as the input. Our previous work on utility-based transcoding [Wang03, Kim03] use a similar architecture to automatically select the optimal frame rate for MC-DCT based video based on the PSNR quality metric. The work in this paper steps further by referring to the subjective evaluation, which makes the result more feasible and persuadable. The experiment result demonstrates that our proposed method can efficiently predict the optimal adaptation matching the subjective evaluation very well.

The rest of this paper is organized as follows. In Section 2, the content-based prediction system is described. The experiment results and analysis are presented in Section 3. Section 4 concludes the paper.

2. Content-based optimal adaptation operation prediction

2.1 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 are 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[Chen03]. 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. We want to find a (or maybe some) adaptation method(s), which output the best quality based on our subjective evaluation criterion.

2.2 Optimal adaptation operation prediction

Our previous work in [Wang03] demonstrated strong correlation between video adaptation behavior and the low level content features. Basically we reasonably assume the videos sharing similar content characteristic will have similar optimal operation preference. During our experiment we have observed in a spatio-temporal adaptation optimization, either spatial or temporal complexity has considerable correlation with the optimum preference. Therefore, we categorize the videos according to their "content complexity". We find the minimum achievable bandwidth ($r_{MAB} \in R$) is a nature partition criterion for video content complexity. r_{MAB} is defined as the bandwidth below which the adaptation operation cannot generate a valid bit stream due to overhead cost coding motion vectors and stream syntax. r_{MAB} well reflects the spatio-temporal integrated content complexity. I.e., we consider the video clips sharing the same r_{MAB} have similar content complexity. The higher r_{MAB} is, the more the complexity is, and vice versa. Given this content complexity categorization, for each category we can summarize a representative operation preference for a given bandwidth considering the statistical characteristic within the category. The representative is then used to guide the adaptation for the incoming video clip.

Therefore, we model the optimal operation problem as a typical pattern classification procedure, where the content features are considered as input observations, while the category as the output label. Thus, our proposed content-based optimal operation prediction system can be illustrated using the diagram in Figure 1. For any incoming video clip, firstly its content features are extracted. These content features, reflecting spatio-temporal complexity of the video, then go through a classification discriminative function and the content category is decided. The classification discriminative function is trained through classic supervised classification routine. Afterwards the optimal preference is recommended and used by the video adaptor. Lastly the adapted video stream is generated.

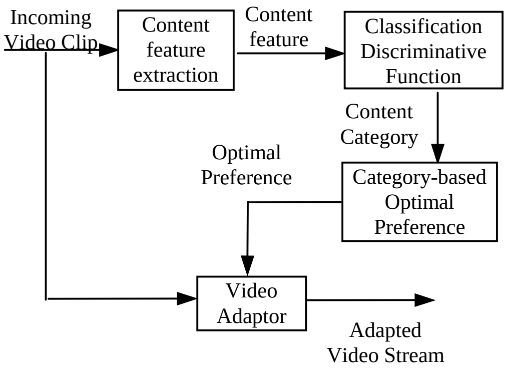

Figure 1: Diagram for content-based optimal adaptation

3. Experiment Result

3.1 Experiment setup

The experiment setup was detailed in [Yong04]. To be brief, we ran a thorough subjective experiment with 128 clips. All video clips were 288-frame long, with CIF resolution and an original frame rate of 30fps. They were coded with the MC-EZBC codec[Chen03] by utilizing three different temporal adaptation approaches: full frame rate (30fps), half frame rate (15fps) and quarter frame rate (7.5fps). The bandwidths tested in the experiment were {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 value of r_{MAB} through adaptation vary. We carried out the subjective evaluation based on the double stimulus impairment scale (DSIS) recommended by ITU-R standard [ITUR00]. Totally 31 subjects participated in the experiment. 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). The subjective scores were collected and analyzed by using statistical testing method and the users' frame rate preferences on each clip and each bandwidth were obtained. Ties are possible and allowed to indicate non-unique preferences. Detailed results about the subjective experiment can be found in [Yong04].

3.2 Optimal operation prediction

3.3.1 Video clip partition

The video clips were partitioned into categories according to their content complexity based on the value of r_{MAB} as introduced in Section 2.2. In our case, r_{MAB} had three distinct values: 50, 100 and 200Kbps, correspondingly the clips were categorized as low, medium and high content complexity. Each category of videos has its representative frame rate preference, which is used during the practical optimal adaptation selection.

3.3.2 Classification performance

The classification discriminative function was trained based on a supervised classification training procedure. In the classificatiovn step, each observation was defined as a subclip with one-GOP length. There were totally 2275 observations. Each subclip used the category label for the video it came from. A cross-validation study was employed during the learning, where 70% of the observations were put in the training pool and the remaining the testing pool. We employ the multiple layer perceptron (MLP) as our classification discriminative function. Other classification models may also be used. The result reported below was averaged over 5 runs.

We adopt the following content features during the classification and prediction:

1) Texture: The texture energy for each of the 5 temporal decomposition levels, which is defined as the summation of squared coefficient magnitudes from all spatial decomposition levels.
2) Motion: The 10-bin motion magnitude histogram for the subclip.
3) BlkHist: the 5-bin block size histogram for the subclip. According to the implementation in [Chen03], during the block-based motion compensation, the block size varied according to the content complexity.

Figure 2 is the classification accuracy result. The performance using each set of content features and all available features are compared together. This result demonstrates that by considering the spatio-temporal content information together, we can accomplish a promising prediction performance (around 81% accuracy) with the proposed system. As a comparison, merely using texture (spatial) or motion (temporal) information cannot fulfill the prediction efficiently, though each has its own contribution. As mentioned above, BlkHist reflects both the texture-motion information in some degree, therefore it achieves a better result than texture or motion does.

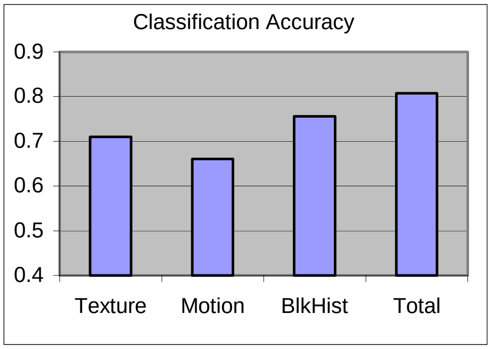

Figure 3: The classification accuracy

3.3.3 Optimal operation prediction accuracy

The classification is not our ultimate purpose. Our destination is to use the prediction result to guide the optimal operation in practice. We estimate the optimal operation prediction accuracy (OOPA) in such a way. Firstly for each video category, a representative optimal operation is obtained by choose the most frequently preferred frame rate within the category. Given a bandwidth r, the prediction accuracy is calculated as:

$$OOPA_r = \frac{Number\ Of\ corret\ prediction\ based\ on\ representative}{Number\ of\ observation}$$

A prediction is considered correct when the representative optimal operation matches (one of, in case of tie) the actual optimal operation(s). Figure 3 is the result of OOPA for each bandwidth. As a comparison, the corresponding result without classification is also estimated, where the representative optimal operation is calculated using the same way mentioned above while over the whole video pool. A notable improvement can be seen, especially at the low bandwidth where our interest focuses on mostly. It is also visible that for different bandwidth, the prediction performance varies, with a sink during the medium bandwidths (200~600kbps). This phenomenon comes from the divisibility characteristic for different bandwidth. Based on the subjective experience, for these bandwidths it is more difficult to specify the optimal operation, resulting a large variance of subjective score. Figure 4 shows the entropy value of the optimal operation distribution for each bandwidth, which in some degree can be considered as the measurement of the prediction difficulty. Actually our proposed method coincides with the entropy analysis very well, while the result without classification cannot.

4. Conclusion

In this paper, make our contribution in the content-based prediction of optimal operation selection in a video adaptation scenario. We use subjective evaluation as our optimal quality criterion and utilize the content-based prediction framework to select optimal spatio-temporal video adaptation operation. Our experiment result indicates that the proposed method can efficiently predict the right adaptation operation matching subjective quality with promising performance.

Reference:

[Chang02] S.F. Chang, "Optimal Video Adaptation and Skimming Using a Utility-Based Framework", IWDC-2002, Capri Island, Italy, Sept. 2002.

[Wang03] 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.

[Kim03] J.-G. Kim, Y. Wang, S.F. Chang, "Content-Adaptive Utility-Based Video Adaptation", in Proc. ICME-2003, Baltimore, Maryland, July 2003. [VQEG00]

[Raj02] R. Kumar Rajendran, M. van der Schaar, S.-F. Chang, FGS+: Optimizing the Joint Spatio-Temporal Video Quality in MPEG-4 Fine Grained Scalable Coding, IEEE International Symposium on Circuits and Systems (ISCAS 2002), Phoenix, Arizona, May 2002.

[Yadavalli03] Yadavalli, G., Masry, M. and Hemami, S.S., Frame Rate Preferences in Low Bit Rate Video, IEEE Intl. Conf. on Image Processing, Barcelona, 2003

[Chen03] Peisong Chen and John W. Woods. "Bidirectional MC-EZBC With Lifting Implementation". Accepted for publish in IEEE trans. on Circuits and Systems for Video Technology, 2003

[Wang04] Yong Wang, Tian-Tsong Ng, Mihaela van der Schaar, Shih-Fu Chang, Predicting Optimal Operation of MC-3DSBC Multi-Dimensional Scalable Video Coding Using Subjective Quality Measurement. Proc. SPIE Video Communications and Image Processing (VCIP), San Jose, CA, January 2004

[Yong04] Yong Wang, Shih-Fu Chang, Alexander C. Loui. Subjective Preference of Spatio-Temporal Rate in Video Adaptation Using Multi-Dimensional Scalable Coding. Submitted to ICME 2004 special session.

[ITUR00] 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.

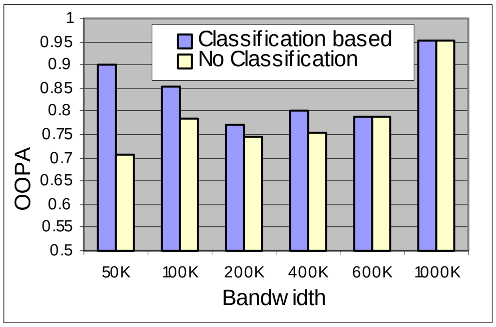

Figure 4: Optimal operation prediction accuracy

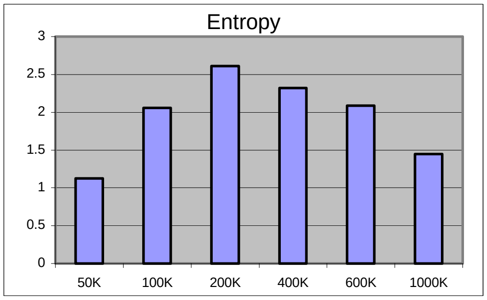

Figure 5: Optimal operation entropy statistic