OPTIMAL RATE AND INPUT FORMAT CONTROL FOR CONTENT AND CONTEXT ADAPTIVE VIDEO STREAMING

Tanır Özçelebi[1], A.Murat Tekalp[1,2], M.Reha Civanlar[1]

[1] College of Engineering, Koc University, Istanbul, Turkey
[2] Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627

ABSTRACT

A novel dynamic programming based technique for optimal selection of input video format and compression rate for video streaming based on "*relevancy*" of the content and user context is presented. The technique uses context dependent content analysis to divide the input video into temporal segments. User selected relevance levels assigned to these segments are used in formulating a constrained optimization problem, which is solved using dynamic programming. The technique minimizes a weighted distortion measure and the initial waiting time for continuous playback under maximum acceptable distortion constraints. Spatial resolution and frame rate of input video and the DCT quantization parameters are used as optimization variables. The technique is applied to encoding of soccer videos using an H.264 [1] encoder. The improvements obtained over a standard H.264 implementation are demonstrated by experimental results.

1. INTRODUCTION

Classical approaches to rate control in video encoding do not take the relevancy of the encoded content into account [2-4]. For example, the Rate-Distortion Optimization (RDO) method [5] minimizes the distortion under a bitrate constraint for the entire content. However, the user utility in a video streaming system can be considerably improved by content-adaptive optimization of the compression parameters. A streaming technique that considers content issues in sports videos is described in [6], where the input video is segmented and encoded in two streams for different relevance levels with "predetermined bitrates," namely, high target bitrate (highly relevant) and low target bitrate (less relevant) streams. The less relevant shots are then encoded such that they are shown as still images at the receiving side and the more important shots are encoded at full quality.

In this paper, we present a new delay-distortion optimization (DDO) approach, where in addition to distortion; the initial waiting time (playback delay) at the receiving side of a video streaming system is minimized considering the relevancy of different parts of the video content to be compressed. Given a particular transmission bitrate, the encoder needs to make tradeoffs between quantization (quality), spatial resolution and frame rate, which are well-adapted to (spatio-) temporal content characteristics of video. Considerable improvements in visual quality and user utility can be achieved for a variety of bitrates and resolutions by this optimization approach.

In section 2, we discuss the quality and relevance measures that are used in this work. Section 3 presents the formulation of our content adaptive delay-distortion optimization (DDO) approach. Experimental results are given in Section 4. Conclusions are drawn in Section 5.

2. PERCEPTUAL QUALITY AND SEMANTIC RELEVANCE MEASURES

An essential part of our formulation is the definition of a perceptual quality measure. Furthermore, "relevancy" of the content also needs to be quantified. Although we present here our particular selections for these measures, our general formulation can be used with any other measure.

2.1. Perceptual Quality Measures

Many perceptual quality metrics have been proposed in the literature [7-9]. Two of the well known metrics are the PSNR and the *blockiness* measures. Insufficient frame rate must also be considered as a source of perceptual disturbance, especially when there is high amount of movement in the clip. Therefore, the overall distortion metric has to consider frame rate along with the blockiness and the PSNR measures.

In our formulation, perceptual video quality and semantic relevance measures are determined at the encoder (server) side, which has access to uncompressed or a high-quality compressed version of the content. Also, PSNR measure has to refer to the original clip. Therefore,

0-7803-8554-3/04/$20.00 ©2004 IEEE.

we use referenced measures, which employ the original video.

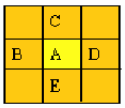

Fig. 1. Organization of blocks in a frame

Blockiness and flatness measures are obtained by a modification to the technique used in [9], which compares pixel intensity variations across boundaries of blocks and within blocks. For example, for MxN blocks, a horizontal blockiness measure, $\mathbf{BM_h}$, between blocks A and B (depicted in Figure 1) for both the original and the encoded versions is computed as follows:

$$BM_h = \begin{cases} \dfrac{BD1_h}{BD3_h}, & if \quad BD3_h \neq 0 \\ 0, & if \quad BD3_h = 0 \end{cases} \tag{1}$$

where, $\mathbf{BD1_h}$ and $\mathbf{BD3_h}$ refer to one-pixel inter-block difference and cumulative difference over ±3 columns across the block boundary, respectively, which are defined by:

$$BD1_h = \gamma_1 \sum_{i=1}^{N} |a_{i1} - b_{iM}| \tag{2}$$

$$BD3_h = \gamma_2 \sum_{i=1}^{N} \left(\sum_{j=M-3}^{M-1} |b_{i(j+1)} - b_{ij}| + \sum_{j=1}^{3} |a_{i(j+1)} - a_{ij}| \right) + \sum_{i=1}^{N} |a_{i1} - b_{iM}| \tag{3}$$

Here $\mathbf{a_{ij}}$ and $\mathbf{b_{ij}}$ denote values of pixels in blocks A and B, respectively, γ_1 and γ_2 are normalization factors. The effective horizontal blockiness of a certain block BM_h^{eff} caused by lossy compression is:

$$BM_h^{eff} = \max\left(\left(BM_h^{enc} - BM_h^{org} \right), 0 \right)$$

where, $\mathbf{BM_h^{enc}}$ and $\mathbf{BM_h^{org}}$ are the horizontal blockiness measures of the same block in the encoded and the original clips, respectively. The effective vertical blockiness measure, $\mathbf{BM_v^{eff}}$, between blocks A and C is defined similarly. Then, the effective blockiness measure for block A is computed as the average of the horizontal and vertical effective blockiness measures between blocks A and B, and A and C; and an overall effective blockiness measure for a frame is defined as the average of the effective blockiness measures of all blocks within that frame. Similarly, a horizontal flatness measure, $\mathbf{F_h}$, between blocks A and B is defined as:

$$F_h = \gamma_3 \sum_{i=1}^{N} \left(\sum_{j=M-3}^{M-1} z(b_{i(j+1)}, b_{ij}) + \sum_{j=1}^{3} z(a_{i(j+1)}, a_{ij}) \right) + \gamma_3 \sum_{i=1}^{N} z(a_{i1}, b_{iM}) \tag{4}$$

where,

$$z(\alpha, \beta) = \begin{cases} 1, & if \quad \alpha = \beta \\ 0, & if \quad \alpha \neq \beta \end{cases} \tag{5}$$

and γ_3 is a normalization factor. A vertical flatness measure, $\mathbf{F_v}$, is computed likewise. The effective flatness measure of a block can be computed by the same procedure used in effective blockiness. Finally, the blocking artifact measure of a certain block is given by, the maximum of the effective blockiness and flatness measures that are appropriately scaled, and the overall blocking artifact measure of a frame is the average of these measures for all blocks that fall within that frame.

2.2 Semantic Relevance and Content Analysis

In almost all video programs, not all the shots are as interesting to a user as others. For example, in a tennis game, breaks given between sets are not as relevant as the in-game strife in most cases. Each of these predefined shot types determine the temporal segments of a specific video type. In [10], a technique for automatic shot classification for video inputs with a specific context such as a soccer game or TV news is presented. After the shots are classified, a user's levels of interest in certain type of shots (points, breaks, commercials etc.) can be accessed prior to the video transmission. These levels of interests are specified between 0 and 1 and are defined as semantic relevance factors (as explained in Section 3). If the user's interests change, the relevance factors can be adjusted accordingly. Different shots that are classified under the same content type will have the same relevance factors.

3. OPTIMAL RATE CONTROL

Our target is to send the relevant video content with maximum perceptual video quality and minimum initial playback delay given the channel bandwidth, and, to never send any content under an acceptable perceptual quality level. For example, when coding sports videos for very low bitrate wireless applications, we would transmit relevant segments at an acceptable video quality, at the expense of dropping less relevant segments, rather than transmitting all segments with less than a minimum acceptable perceptual quality.

We assume that the video consists of **N** temporal segments and each segment is encoded by a set of spatial and temporal resolutions and different sets of quantization parameters. We formulate the selection of the best encoding parameters for each segment as a function of its relevance and perceptual distortion, given fixed channel bandwidth (assuming channel is lossless), minimum initial delay at the receiver and quality constraints as an off-line utility-resource optimization problem. The quantization step sizes for both the intra and inter coded frames are determined as in [2]. Assuming that each of the **N** segments, with semantic relevance factors $\{w_1, w_2, \ldots, w_N\}$, has been coded off-line using **k** combinations of spatial resolutions, frame rates, and quantization parameters, we

store the perceptual distortion measures achieved for each segment:

$\{D_1^1, D_1^2, \ldots, D_1^k, D_2^1, D_2^2, \ldots, D_2^k, \ldots, D_N^1, D_N^2, \ldots, D_N^k\}$

where, each $\mathbf{D_i^j}$ is a weighted sum of the blockiness, PSNR and the jitter measures (increasing PSNR has a negative effect on distortion). The jitter measure due to insufficient frame rate is computed as the difference of average motion vector lengths between full frame rate and the current frame rate.

Bitrates corresponding to the above distortions:

$\{R_1^1, R_1^2, \ldots, R_1^k, R_2^1, R_2^2, \ldots, R_2^k, \ldots, R_N^1, R_N^2, \ldots, R_N^k\}$

are also stored for each combination of these encoding parameters. The optimal set of encoding parameters for each segment is then chosen by solving a constrained, multi objective optimization problem to minimize the initial playback delay and the weighted distortion at the receiver subject to maximum acceptable distortion constraints $\mathbf{D_i^{max}}$ and a "*continuous playback*" constraint, guaranteeing non-stop playback at the receiver if the transmission bandwidth stays at or above the specified rate for the duration of the transmission. This formulation can be stated as:

$$\min_j(t_w) = \min_j\left\{\sum_{i=1}^{N}\frac{R_i^{\,j} - BW}{BW} y_i^j \cdot TD_i\right\} \qquad (6)$$

$$\min_j(D) = \min_j\left\{\sum_{i=1}^{N} w_i \cdot D_i^{\,j} \cdot y_i^j \cdot TD_i\right\} \qquad (7)$$

jointly subject to

$$D_i^{\,j} \le D_i^{\,\max}, \quad i = 1,\ldots,N$$

and

$$t_w \cdot BW - \sum_{i=1}^{n} y_i^j (R_i^{\,j} - BW) TD_i \ge 0, \quad n = 1,\ldots,N$$

where, $\mathbf{R_i^j}$, $\mathbf{TD_i}$ and **BW** are the bitrate, the duration of the $\mathbf{i^{th}}$ video segment and the available bandwidth of the channel respectively, and $\mathbf{y_i^j}$ is a binary variable denoting if the specific shot is actually encoded for transmission ($\mathbf{y_i^j}$ = 1) or skipped ($\mathbf{y_i^j}$ = 0). Here, the variable $\mathbf{R_i^j}$ is a function of the coding parameters, that is, the quantization step-size, frame rate and spatial resolution. The minimization is over the value of **j=1,…,k** for each temporal segment **i**. The last constraint guaranties that we never stop streaming after an initial waiting time.

One of the well known solution techniques for **multi objective** dynamic programming problems as the one above is finding an optimal *utopia* point for each of the objective functions and, then, finding the best compromise by examining all feasible points in between these utopia points. Software packages exist for the solution of such problems. In our study, we used *General Algebraic Modeling System (GAMS) Integrated Development Environment*[1] software.

4. EXPERIMENTS AND RESULTS

For our experiment, we chose a 16.12 seconds long soccer video, which is divided into 4 different shots using the content analysis technique of [10]. The first shot is a goal possession that is of great interest to most users, the second shot is a scene where the players cuddle to celebrate after the goal. The audience is shown on the third shot and finally, the team coach is seen on the last shot. The precise relevance weights of these shots depend on the user's choice. For example, if the user doesn't want to see the parts where the audience is shown, the weight of that shot should be zero. In this case, the optimal encoding result may not include this irrelevant shot at all.

Shot	Weight	Resolution	FPS	Bitrate	Duration
1	1	176x144	7.5	108.21 kbps	4 sec
2	0.25	96x80	7.5	34.32 kbps	7.76 sec
3	0.125	96x80	7.5	38.99 kbps	2.3 sec
4	0.125	176x144	7.5	55.62 kbps	2.06 sec

Table 1: Optimal parameters for the video sample

Table 1 shows the optimal selections for the relevancy factors (weights) used for this example. Figure 2 shows a comparison of frames from different types of shots taken from content adaptive coded version and the regular version (JM 7.4 from the JVT group) of the same QCIF soccer clip encoded at the same rate. The clip is coded at an overall bitrate of 56.04 kbps and the channel bandwidth is assumed to be 25 kbps, resulting in an overall waiting time of 20.02 seconds for the content adaptive case at the receiving side. For the standard case, the waiting time becomes 20.82 seconds. Note that since the first segment of this example is a high-relevancy one, the gain in the start up delay is not that significant. While the ball and the lines of the field are quite noticeable in the content adaptive coded clip, we can't see the ball and certain parts of the pitch lines in the regular encoded version. Also, for the 2nd shot, the blocking artifacts are very distracting in this version.

5. CONCLUSIONS

In this paper, we introduce a new delay-distortion optimization approach to content-based rate control using dynamic programming. Using semantic relevancy in the determination of encoding parameters is found to be very efficient for streaming videos over small bandwidth channels. This method not only maximizes perceptual quality of relevant parts in the video (e.g. the long shots in Figure 2), but also minimizes the playback delay at the receiving side. Our algorithm outperforms the regular rate control in the visual quality of the long shots and still retains an acceptable quality in other shots. It is expected that, DDO would perform better irrespective of which coding technology it is built upon since it saves bandwidth

for high relevance parts while transmitting low relevance segments.

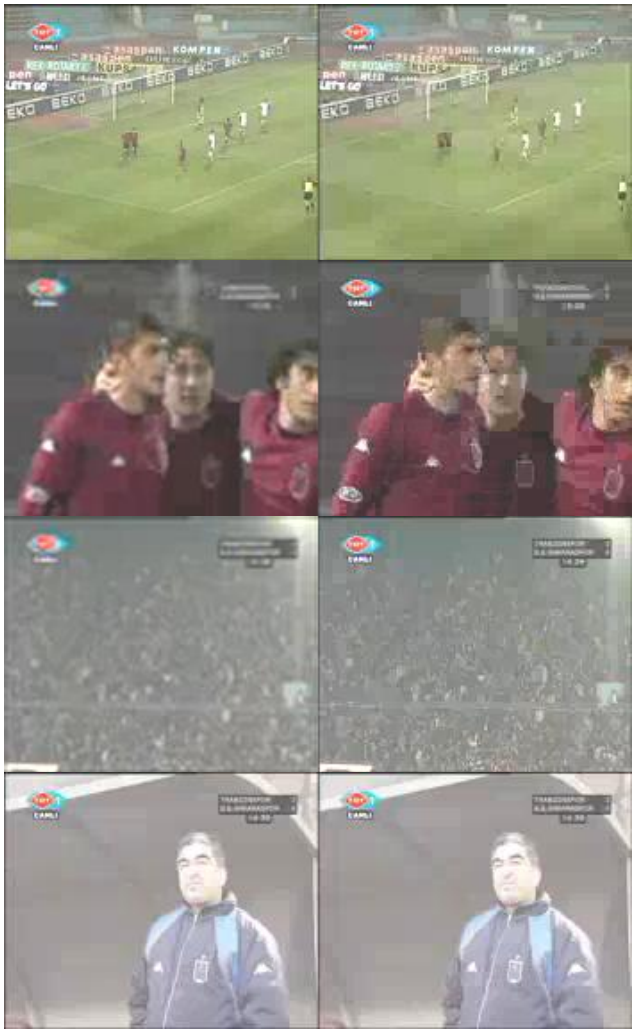

Fig. 2. Shots taken from content adaptively coded (on the left) and the RDO optimized coded (on the right) clips at 56.04 kbps.

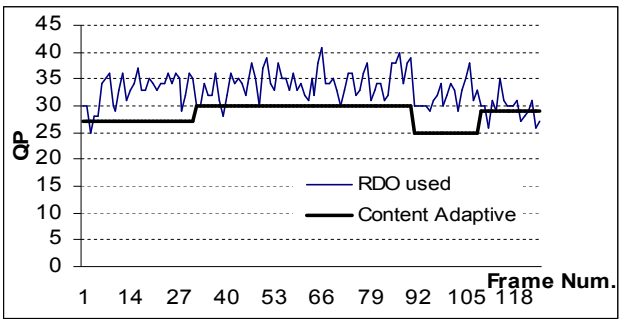

Fig. 3. Quantization Parameters used in each frame

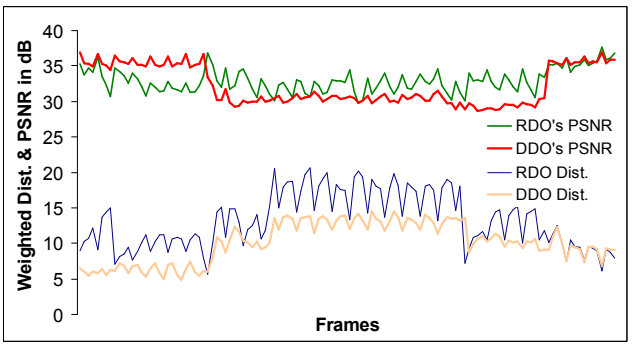

Fig. 4. PSNR and weighted distortion of individual frames

6. REFERENCES

[1] T. Wiegand, G. Sullivan and A. Luthra, "Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC)," May 27, 2003.

[2] S. Ma, W. Gao, F. Wu, Y. Lu, "Rate Control for JVT Video Coding Scheme with HRD Considerations," ICIP 2003.

[3] S.Ma, W. Gao and Y. Lu, "Rate Control on JVT Standard," Doc. JVT_D030, Klagenfurt, Austria, 22-26 Jul. 2002.

[4] T. Chiang, Y. Zhang, "A New Rate Control Scheme Using Quadratic Rate Distortion Model," IEEE Trans. Circ. Syst. Video Technology Vol. 7, pp. 287-311, Apr. 1997.

[5] G.J. Sullivan and T. Wiegand, "Rate-Distortion Optimization for Video Compression," IEEE Signal Processing Mag., vol. 15, no. 6, pp. 74-90, Nov. 1998.

[6] S.F. Chang, D. Zhong, and R. Kumar, "Real-Time Content-Based Adaptive Streaming of Sports Video", IEEE Workshop on Content-Based Access to Video/Image Library, pp.139-146, Hawaii, Dec. 2001.

[7] S. Winkler, A. Sharma, D. McNally, "Video Quality and Blockiness Metrics for Multimedia Streaming Applications" in Proc. of the Int. Symp. on Wireless Personal Multimedia Communications, pp. 547-552, Aalborg, Denmark, Sept. 12, 2001.

[8] C.J. van den Branden Lambrecht, "Perceptual Quality Measure Using a Spatio-temporal Model of the Human Visual System" Proceedings of the SPIE, vol. 2668, pp. 450-461, San Jose, 1996.

[9] F. Pan, X. Lin, S. Rahadja, W. Lin, E. Ong, S. Yao, Z. Lu and X. Yang,"A Locally Adaptive Algorithm for Measuring Blocking Artifacts in Images and Videos," submitted to Signal Processing: Image Comm., October 2003.

[10] A. Ekin, A. M. Tekalp and R. Mehrotra, "Automatic Soccer Video Analysis and Summarization," IEEE Trans. on Image Processing, vol. 12, no. 7, pp. 796-807, June 2003.

[1] GAMS Development Corporation, http://www.gams.com