Delay-Distortion Optimization for Content-Adaptive Video Streaming

Tanir Ozcelebi, A. Murat Tekalp, Fellow, IEEE, and M. Reha Civanlar, Fellow, IEEE

Abstract—We propose a new pre-roll delay-distortion optimization (DDO) framework that allows determination of the minimum pre-roll delay and distortion while ensuring continuous playback for on-demand content-adaptive video streaming over limited bitrate networks. The input video is first divided into temporal segments, which are assigned a relevance weight and a maximum distortion level, called relevance-distortion policy, which may be specified by the user. The system then encodes the input video according to the specified relevance-distortion policy, whereby the optimal spatial and temporal resolutions and quantization parameters, also called encoding parameters, are selected for each temporal segment. The optimal encoding parameters are computed using a novel, multi-objective optimization formulation, where a relevance weighted distortion measure and pre-roll delay are jointly minimized under maximum allowable buffer size, continuous playback, and maximum allowable distortion constraints. The performance of the system has been demonstrated for on-demand streaming of soccer videos with substantial improvement in the weighted distortion without any increase in pre-roll delay over a very low-bitrate network using AVC/H.264 encoding.

Index Terms—Content adaptive video coding, video streaming, pre-roll delay, GoP rate allocation, semantic relevance.

I. INTRODUCTION

PRE-ROLL delay is a vital parameter in video streaming since it provides some level of protection against network throughput variations, as well as allowing flexible rate allocation in video coding. If it is chosen too small, pauses in video playback due to network throughput variations and/or unacceptable video quality due to strict rate control in video coding would result. An unnecessarily large pre-roll delay, which in the limit leads to the download-and-play solution, requires a very long initial wait, thus eliminating the benefit of streaming, and is usually found objectionable by users. Therefore, video streaming applications should strike the right balance between pre-roll delay and video distortion. This issue becomes even more significant in content-adaptive video streaming over low-bitrate networks, where different bitrates (sometimes larger than the network throughput) shall be allocated to different temporal video segments (shots) according to their importance.

Streaming video over low-bitrate networks, such as 3G and beyond wireless systems, remains to be a challenging problem even with Quality of Service (QoS) support. Content adaptive video coding has been introduced as a potential solution to this problem [1], where the video is parsed into semantic temporal segments. Important temporal segments are encoded at a high enough bitrate, while the rest is transmitted at a very low bitrate (e.g., as key frames and audio). However, in this early work, the low and high bitrates are determined according to client buffer size and channel bandwidth in an ad-hoc manner. There also exist a number of content-adaptive transcoding strategies: Content-adaptive multimedia access technologies that support Universal Multimedia Access (UMA) are presented in [2] and [3]. In [2], assuming that each spatial region of interest \( R_i \) of a video segment has an importance hint \( 0 \leq I_i \leq 1 \), and a spatial resolution hint \( 0 \leq S_i \leq 1 \), the optimization problem is formulated as finding a set of regions \( R_i \) and a rescaling factor \( L \) such that the overall fidelity score of the rescaled set is maximized and the minimum bounding rectangle surrounding the cropped and rescaled set \( R \) fits the screen size of the receiving device. A method, where transcoding policies are determined by the content author is described in [4]. Depending on the client capabilities, versions of content with various resolutions and modalities are produced off-line, and the version that maximizes a subjective measure of fidelity is selected.

In [5], new performance measures for semantic adaptation, namely Viewing Quality Loss and Bitrate Cost Increase, are discussed. Object or event based segments of the input video are automatically classified into relevance levels. The unequal bit allocation strategy between important and unimportant temporal segments is determined by the semantic statistics (size and number of relevant and irrelevant segments) of the input video and the target bitrate. If a relevant segment is misclassified, a loss of quality occurs and is denoted by Viewing Quality Loss. Conversely, if an unimportant segment is misclassified, an unnecessarily high bitrate will be used, referred as Bitrate Cost Increase.

We recognize that some content adaptive techniques; including the one proposed here, yield temporal variations in quality, which may be unacceptable for entertainment-quality video. On the other hand, over very low bandwidth networks, if such techniques are not used, then almost no valuable visual information may be delivered. For example, when a soccer video is encoded at low bitrates with uniform quality, there may be severe distortions to the extent that the ball and the
players are not visible and pitch lines are lost in the most important scenes (e.g., goals). Content adaptive coding facilitates best effort transmission of such relevant information instead of enforcing an average and low quality for the entire video segment. This paper addresses optimal bit allocation between different temporal segments to minimize distortion and pre-roll delay under pre-set quality-level and continuous playback constraints. An alternative approach was introduced in [6], where rate-distortion optimized video summarization and transmission over packet lossy networks with minimum video distortion has been studied.

The classical approach to ensure continuous video playback for a fixed target encoding rate relies on the buffer management strategy of the underlying codec system, and determines the pre-roll delay as a function of the decoder buffer size (a hardware constraint). For example, the Video Buffer Verifier (VBV) model [7] of MPEG and the Hypothetical Reference Decoder (HRD) model [8] in AVC/H.264 [9]-[10] verify that the bitstream generated by an encoder can be played-back continuously at the decoder given the decoder buffer size and pre-roll delay for a constant bitrate (CBR) channel with a specified rate. However, the effects of the pre-roll delay or the decoder buffer size on the overall distortion are not specified in these models. With software decoders for streaming applications, hardware constraints become less important while the pre-roll delay becomes a main performance parameter (which then determines the required buffer size). In [11], an adaptive media playout (AMP) scheme was proposed as a means to ensure continuous playback, where the client device can adaptively change playout speed of the content in order to prevent buffer overflow and underflow. In [12], AMP framework is combined with the well-known rate-distortion optimized (RDO) [13] streaming. Although AMP addresses continuous playback issue in an ad-hoc manner, in low-bitrate streaming applications with non-uniform bitrate allocation among temporal segments, optimum determination of pre-roll delay under continuous playback constraint remains as an important concern.

In content-adaptive video coding and transcoding, temporal shot detection and relevance assignment methods may have significant effect on the overall performance. Most effective methods are highly context (domain) dependent. For example, in the context of a soccer game, the temporal video segments showing a goal event and the spatial segments around the ball are more important than any other part of the video. In a tennis game, breaks given between sets are not as relevant as the in-game strife. Television news reports can be segmented as anchorperson shots, news footage and commercial breaks. For movies, temporal shot detection and content analysis may facilitate bit rate assignment as a function of coding difficulty and existence of special effects. There exist several techniques in the literature for automatically analyzing such content [14]-[18]. Automatic content analysis is beyond the scope of this paper, and we assume appropriate content analysis tools are available.

Another essential part of our framework is the definition of distortion and semantic relevance measures for video content. Although Peak Signal-to-Noise Ratio (PSNR) is the most commonly accepted distortion metric in the literature, it is not always a good indicator of perceptual quality when spatial and temporal resolutions are varied in rate allocation; hence the need for richer quality metrics [19]-[20]. Blockiness and flatness measures have been more closely linked with perceptual quality. Insufficient frame rate due to frame skipping can also be considered as a source of perceptual disturbance, especially when there is high motion in the clip. Several other perceptual quality metrics have been proposed in the literature [21]-[25]. It is not the objective of this work to develop new video quality metrics, but rather to employ recently published such measures in our problem formulation.

This paper offers the following main contributions: i) A new delay-distortion optimization (DDO) framework for content-adaptive video streaming using multiple-objective optimization (MOO), which allows studying trade-offs between pre-roll delay and distortion is proposed in Section II. ii) A new off-line content-adaptive streaming solution for video-on-demand using this framework, where the best trade-off between spatial and temporal video resolutions (for encoding), and encoder quantization parameters for delay-distortion optimization is provided in Section III. The method proposed in this paper is an off-line procedure for rate allocation to each temporal segment which is applicable to finite length video clips. The main application is on-demand video streaming over limited bandwidth networks with QoS where acceptable video quality must be delivered with minimum delay. First, we encode each temporal segment (also referred as GoP) individually with multiple target bitrates. Rate-distortion optimization (RDO) is used while encoding each segment [7]-[9]. Our proposed solution determines the target rate, and spatial and temporal resolutions for each GoP to achieve the least overall distortion and pre-roll delay for the video according to a user specific relevance-distortion policy, given the temporal segment boundaries. Finally, selected bitstreams for each GoP are pasted together using a bitstream assembly unit. The proposed framework can be used with any video codec, including the state of the art AVC/H.264 encoder.

The paper is organized as follows: Section II discusses the problem formulation for off-line delay-distortion optimized content-adaptive streaming. Section III presents a particular linear programming (LP) solution. Section IV presents experimental results, where we observe considerable improvements in visual quality and user utility for a variety of bitrates using our bit allocation approach. In Section V, conclusions are drawn. Finally, the Appendix overviews the main principle of the MOO approach which is used in our solution.

II. PROBLEM FORMULATION

In this work, we assume that a video clip has already been partitioned into N temporal segments. Our goal is to send more relevant temporal segments with high perceptual quality and minimum pre-roll delay over a CBR channel with bitrate
given a specific relevance-distortion policy, and never to send any content under an acceptable perceptual quality level. Clearly an acceptable quality (lower distortion) can be attained by increasing pre-roll delay (encoding at a rate higher than $R_\text{ch}$). We first introduce the relevance-distortion policy for content-adaptive video streaming in Section II.A. Section II.B addresses the relationship between pre-roll delay and distortion for the case of variable target bitrates for each segment, under continuous playback constraint. We then formulate selection of the best encoding parameters for each segment as a multiple objective optimization problem, to minimize the perceptual coding distortion and pre-roll delay at the receiver in Section II.C, where maximum buffer size, continuous playback and the maximum perceptual distortion (per segment) constraints are taken into account.

A. Relevance-Distortion Policy

A relevance-distortion policy assigns a relevance level $w_n$ and a maximum allowed distortion $D_n^{\text{max}}$ for each temporal segment $n$ according to its content. This policy may be universal, set at the server side, or may be user-specific, provided in a user profile. In applications where not all temporal segments may be equally interesting to a user, relevance levels can depend on the semantics of the content; in other contexts, relevance levels may be assigned according to low level features or coding difficulty. That is, a user’s level of interest in certain shots (e.g., goals in soccer, touch-downs in football) in a given video context can either be set to default values derived from general opinions, or it can be signaled by a specific user prior to streaming session. It is also possible to compute relevance level automatically from audio and video features [26]-[27]. For example, in a sports video, if we assume that audio signal energy increases whenever an important event occurs since the voice of the commentator and/or the noise of the audience will increase, then the relevance level of video segment $n$ may be defined by

$$w_n = \frac{E_n}{E_\text{global}}$$

where $E_n$ denotes the average audio energy of segment $n$, and $E_\text{global}$ is the average audio energy for entire video. The relevance factors are normalized between 0 and 1. We note that our framework does not depend on any specific method of determining the relevance factors.

It is important to specify a suitable distortion measure for video. This measure can be PSNR, perceptual quality measures, or a combination of both. We specify maximum allowed distortion levels $D_n^{\text{max}}$ for each video segment, as a function of the relevance of each segment, such that we would not transmit a video segment at a quality less than this specified level for an acceptable video experience.

B. Delay-Distortion Optimization with Continuous Playback Constraint

In this section, we consider how to ensure continuous playback in content-adaptive video encoding/streaming, where different target bitrates, $R_1, \ldots, R_N$, will be assigned to different temporal segments of the video. Assume that the duration of video, $TD$ sec., has been divided into $N$ temporal segments, and that the target bitrate $R_n$ for each segment $n$ is fixed. The minimum buffer size $B_n$ to account for within-segment bitrate variations is an input to the reference decoder buffer verifier model of the particular codec that we use to code the corresponding segment. Hence, for continuous playback, the pre-roll delay must be chosen to guarantee that the receiver buffer must have at least $B_n$ bits at the start of each segment $n$ for the entire duration. Therefore, a necessary condition for $T_{\text{pre}}$ to satisfy

$$R_{\text{ch}} \cdot T_{\text{pre}} + R_n \cdot \sum_{i=1}^{n} TD_i \geq \sum_{i=1}^{n} R_i \cdot TD_i + B_{n+1} \text{ for all } 0 \leq n \leq N$$

where $TD_i$ denotes the duration of segment $i$ and $B_{N+1} = 0$. The first term on the left hand side is the number of bits accumulated in the decoder buffer during the pre-roll period, the second term is the total bits received until playback of segment $n$ is complete, the first term on the right hand side is the total bits drawn from the decoder buffer for playback of first $n$ segments, and the second term is the number of bits that must be present in the buffer before the start of segment $n+1$ to make sure continuous playback during segment $n+1$ according to the reference decoder buffer verifier model of the particular codec. Therefore, a necessary condition for continuous playback for the whole video can be stated as:

$$T_{\text{pre}} \geq \max_{0 \leq n \leq N} \left\{ \frac{\sum_{i=1}^{n} R_i \cdot TD_i + B_{n+1}}{R_{\text{ch}}} \right\}$$

$$= \max_{0 \leq n \leq N} \left\{ \sum_{i=1}^{n} TD_i \left( \frac{R_i}{R_{\text{ch}}} - 1 \right) + \frac{B_{n+1}}{R_{\text{ch}}} \right\}$$

Observe that the value of $T_{\text{pre}}$ to ensure continuous playback depends on how target bitrates are assigned to different temporal segments, hence the given relevance-distortion policy, although the average bitrate and duration of the clip are the same. This is demonstrated by a simple example below.

Example: A video clip, with duration $TD$ and $N=2$ segments, shall be encoded in two different ways:

a) First segment, with duration $TD_1 = \frac{2}{3} TD$ is encoded at $R_1 = 96$ kbps and second segment $TD_2 = \frac{1}{3} TD$ at $R_2 = 32$ kbps;

b) First segment, with duration $TD_1 = \frac{1}{3} TD$ is encoded at $R_1 = 32$ kbps and second segment $TD_2 = \frac{2}{3} TD$ at $R_2 = 96$ kbps, as depicted in Fig. 1. The average bitrate for both cases is the same (74.67 kbps). Assuming the channel bitrate is $R_{\text{ch}} = 64$ kbps, let us now calculate $T_{\text{pre}}$ required for continuous playback for each case.

$\sum$ Summations are assumed to be zero when the lower index is larger than the upper index.
Case a) The minimum pre-roll delay is given by
\[
T_{\text{pre}} \geq \max \left\{ \frac{B_1}{64} \frac{2}{3} TD \left( \frac{96}{64} - 1 \right) + \frac{B_2}{64} \frac{2}{3} TD \left( \frac{96}{64} - 1 \right) + \frac{1}{3} TD \left( \frac{32}{64} - 1 \right) \right\}
\]
\[
= \max \left\{ \frac{B_1}{64} \frac{1}{3} TD + \frac{B_2}{64} \frac{1}{6} TD \right\} = \max \left\{ \frac{B_1}{64} \frac{1}{3} TD + \frac{B_2}{64} \right\}
\]
Case b) The minimum pre-roll delay is given by
\[
T_{\text{pre}} \geq \max \left\{ \frac{B'_1}{64} \frac{1}{3} TD \left( \frac{32}{64} - 1 \right) + \frac{B'_2}{64} \frac{1}{3} TD \left( \frac{32}{64} - 1 \right) + \frac{2}{3} TD \left( \frac{96}{64} - 1 \right) \right\}
\]
\[
= \max \left\{ \frac{B'_1}{64} \frac{1}{3} TD + \frac{B'_2}{64} \frac{1}{6} TD \right\}
\]

We observe that the required minimum pre-roll delay can differ depending on how rate is allocated to each segment even though the average encoding rates and channel conditions are the same. In this setup, the pre-roll delay for case (a) could be more than twice the pre-roll delay for case (b) depending on the values of \(B_1, B'_1, B_2, B'_2\). We note that these values will depend on the coding pattern (IBBBPBBBP...) and encoding parameters used for the temporal segments.

Hence, in content-adaptive (variable target bitrates for segments) video streaming systems, there exists a trade-off between pre-roll delay and relevance-distortion policy used, similar to the well-known rate-distortion trade-off in fixed target bitrate encoding/streaming systems. In streaming applications with segment-based content-dependent target bitrates; however, applying classical RDO solution to each segment with different target bitrates does not necessarily guarantee the best overall visual experience and minimum pre-roll delay for continuous playback. It is well known that larger values of \(T_{\text{pre}}\) provide more flexibility in assigning higher rates to particular segments, as well as more latitude in the allocation of rates \(R_1, ..., R_N\) to each segment, hence a better visual experience at the expense of a larger buffer requirement and initial wait time. Therefore, we propose a delay-distortion optimization (DDO) formulation in the following to strike a compromise between pre-roll delay and overall distortion to obtain the best pre-roll delay vs. distortion performance.

**C. Multi-Objective Optimization Formulation**

Given a relevance-distortion policy, we propose a multi-objective optimization formulation for delay-distortion optimization, where the optimal encoding parameters, hence the rates \(R_1, ..., R_N\) for each segment are determined to minimize the pre-roll delay and weighted overall distortion, \(D\), at the receiver subject to maximum acceptable average distortion \(D_n^\text{max}\) for each segment \(n\) and a maximum buffer size constraint. That is,

\[
\min \left\{ T_{\text{pre}} \right\} = \min \left\{ \max \{R_n^\text{opt} \} \sum_{i=1}^N \left( \frac{R_i}{R_{ch}^i} - 1 \right) + \frac{B_{n,\text{ext}}}{R_{ch}^i} \right\}
\]

\[
\min \left\{ D \right\} = \min \left\{ \sum_{n=1}^N w_n \cdot D_n \cdot TD_n \right\}
\]

jointly subject to
\[
D_n \leq D_n^\text{max}, \quad n = 1, ..., N
\]

and
\[
B_{n+1} \leq R_{ch} \cdot T_{\text{pre}} + R_{ch} \cdot \sum_{i=1}^n TD_i - \sum_{i=1}^n R_i \cdot TD_i \leq B_{\text{max}}
\]

for all \(n = 0, ..., N\), where \(D_n\) and \(w_n\) denote the average distortion and relevance measure for temporal segment \(n\) respectively, and \(B_{\text{max}}\) is the maximum buffer size at the decoder. Minimization is performed over values of \(R_n\) for each temporal segment \(n\).

The objective function in Eqn. 3 is derived from the continuous playback constraint in variable target bitrate scenario explained in Section II.B and aims to minimize the initial wait time. The constraint given by Eqn. 6 denotes the necessary condition to guarantee continuous playout, and it imposes that there is no buffer overflow or underflow at shot boundaries. We make the following observations and note that the proposed formulation includes some well-known solutions as special cases.

1) If \(D_n^\text{max}\) constraints (Eqn. 5) are not used, then distortion of a particular segment can be unacceptable. For example, the ball or field lines may be distorted in low-bitrate sports streaming.

2) If buffer size constraint (Eqn. 6) is not used, arbitrary \(D_n^\text{max}\) constraints can be satisfied at the expense of increased pre-roll delay \(T_{\text{pre}}\) by encoding at a rate higher than the channel rate \(R_{ch}\).

3) If objective \(T_{\text{pre}}\) (Eqn. 3) is not minimized, then the optimal solution approaches the download and play solution.

4) If objective \(D\) (Eqn. 4) is not minimized, then it may result in underutilization of the channel bandwidth when the minimum value of \(T_{\text{pre}}\) is zero, with the trivial solution such that \(D_n = D_n^\text{max}\), for all \(n\) where each segment is encoded with the worst allowable distortion. The multi-objective optimization solution allows allocation of the excess rate in certain segments to achieve a smaller distortion in the future segments.

5) It is not possible to simply minimize the average rate subject to distortion constraints (Eqn. 6) and achieve the minimum pre-roll delay. See the example in Section II.B.

6) If no feasible solution exists, because the conflicting maximum distortion \(D_n^\text{max}\) (Eqn. 5) and maximum buffer size \(B_{\text{max}}\) (Eqn. 6) constraints cannot be satisfied simultaneously, then we try discarding the segment with the least relevance value and/or shortest duration, and try again.

**III. AN OFF-LINE DELAY-DISTORTION OPTIMIZATION SOLUTION**

In this section, we provide a particular off-line solution to the delay-distortion optimization problem formulated in Section II using the AVC/H.264 video codec.
A. Linear Programming Solution

In our solution, the rates \( \{R_1, \ldots, R_N\} \) will be indirectly determined as a function of a set of encoding parameters, the frame rate (temporal resolution), picture size (spatial resolution), and quantization parameter (SNR resolution), which are the independent optimization variables for each segment.

We assume that the frame rate, picture size and quantization parameter for each segment is quantized to certain predetermined levels for a total of \( K \) possible combinations. Each of the \( N \) segments, with semantic relevance factors \( \{w_1, w_2, \ldots, w_K\} \), has been coded off-line using these \( K \) combinations of spatial resolutions, frame rates, and quantization parameters. In our study, the PSNR and blockiness measures are computed in comparison to the original video at the highest spatial resolution after spatial interpolation of the encoded-decoded video as needed. The average perceptual distortion measures for each segment are \( \{D_1^1, D_1^2, \ldots, D_1^K, D_2^1, D_2^2, \ldots, D_2^K, \ldots, D_N^1, D_N^2, \ldots, D_N^K\} \), where the subscript denotes the segment count and the superscript denotes a particular combination of coding parameters. Each \( D_n^k \) has been calculated as a weighted sum of PSNR and blockiness measures (increasing PSNR has a negative effect on distortion) given by

\[
D_n^k = \frac{\text{Blk}_{\text{max}} - \text{Blk}_{\text{min}}}{\text{PSNR}_{\text{max}} - \text{PSNR}_{\text{min}}} \cdot \text{PSNR}_{\text{min}} - \text{PSNR}_{\text{min}}
\]

where \( \text{Blk}_{\text{min}}, \text{Blk}_{\text{max}}, \text{PSNR}_{\text{min}} \) and \( \text{PSNR}_{\text{max}} \) denote the minimum and the maximum of blockiness and PSNR measures [21], achieved respectively, computed over all shots. A motion jitter measure to account for insufficient frame rate, if included, can be computed as the difference of average frame rates between full frame rate and the current frame rate. Bitrates corresponding to the above distortions;

\[
\{R_1^k, R_1^k, \ldots, R_1^k, R_2^k, R_2^k, \ldots, R_2^k, \ldots, R_N^k, R_N^k\}
\]

are also computed for each combination of these encoding parameters. The quantization step sizes for both the intra and inter coded frames are determined as in [8]. The resulting \( \{R_n^k, D_n^k\} \) pairs for each coding parameter set \( k \) and segment \( n \) are depicted in Fig. 2.

If the original video is pre-processed to change its spatial and temporal resolution, the distortion measures outlined above become functions of spatial and temporal resolutions selected for the video segment to be encoded as well as the quantization parameters. Hence, selection of the optimal distortion implicitly selects the best spatial and temporal resolution to be used, in addition to the optimal quantization parameter. Therefore, the problem of finding the optimal set of encoding parameters for each segment is then equivalent to finding a particular path on the coding parameter set index versus segment index graph shown in Fig. 2, such that Eqn. 3 and Eqn. 4 are minimized subject to Eqn. 5 and Eqn. 6. Each feasible path in Fig. 2 yields a pre-roll delay and overall distortion pair \( (T_{\text{pre}}, D) \), which corresponds to a point on the two-dimensional delay-distortion graph depicted in Fig. 3.

To find the optimal path, we first determine the utopia point (see Appendix), which is defined as the delay-distortion point obtained by optimizing each objective function individually while ignoring the other. More specifically, we first ignore the delay objective function (Eqn. 3) and find the solution that gives the minimum distortion. This returns the encoding parameter set that yields the point \( \text{Opt}_1=(T_{\text{max}}, D_u) \) in Fig. 3. Next, we ignore the distortion objective function (Eqn. 4) and find the encoding parameter set that gives the minimum pre-roll delay, hence the point \( \text{Opt}_2=(T_{\text{u}}, D_{\text{max}}) \) shown in Fig. 3. The point \( U=(T_u, D_u) \) is called the utopia point.

Next, we determine the set of Pareto-optimal solution points. A delay-distortion pair \( (T_{\text{pre}}, D) \) is called a Pareto-optimal solution if the value of the distortion cannot be decreased without increasing the value of pre-roll delay, and vice versa. The set of Pareto-optimal points is shown by the curve in Fig. 3. In order to find a set of Pareto-optimal solution points, the horizontal axis is uniformly quantized in the interval \( [T_u, T_{\text{max}}] \) using \( Q_T \) levels, and minimum distortion values for the quantized pre-roll delay values are determined using linear programming, where each quantized pre-roll delay is used as an upper bound constraint, disregarding the delay objective function (Eqn. 3). The best compromise solution can only be determined after finding all such constrained solutions and forming the Pareto-optimal curve. Alternatively, it is possible to quantize the distortion axis using \( Q_D \) levels and find the minimum pre-roll delays for \( Q_D \) distortion constrained optimization problems. Software packages exist for solving such linear programming problems. In our study, we used General Algebraic Modeling System.
Finally, the best compromise (optimal) path, hence the set of encoding parameters for each segment, is chosen as the path that corresponds to the closest solution to the utopia point, $U=(T_n, D_n)$, among all Pareto-optimal solutions using a suitable distance measure. An example MOO problem and its solution have been demonstrated in the Appendix.

It is well-known that an LP problem can be solved in polynomial time using optimization methods in the literature such as the projective method $[28]$. In order to find the Pareto-optimal curve, we need to apply the LP procedure $Q_T$ times (the number of quantization levels on the pre-roll delay-axis). Therefore, the computational complexity of the optimization process is $Q_T$ times that of the LP procedure, which is polynomial time in the number of temporal segments $N$. For example; this computation takes approximately 20 minutes on a 3.00 GHz Pentium when $N=1080$ and $Q_T=6$ for a 90 minutes soccer game.

**B. Overall System Summary**

The operation of the proposed encoder and decoder is shown in Fig. 4. The content analysis and shot classification module performs shot boundary detection and classification of each shot into certain pre-defined semantic content types. The output of the module is $N$ temporal segments each with a relevance measure, $w_i$, $n=1,\ldots,N$. The pre-processor converts each segment into pre-selected spatial and temporal resolution format choices. The standard encoder encodes each input segment $I_n$ with all possible encoding parameter sets ($K$ spatial/temporal resolution and quantization parameter choices) resulting in $K\times N$ output segments. The output of the standard encoder for the $j^{th}$ segment and $j^{th}$ encoding parameter set is a bitstream with rate-distortion pair ($R_j^i, D_j^i$). After this stage, all rate-distortion pairs for each temporal segment along with user-defined relevancy levels and available channel bandwidth information are fed to the MOO module. The optimal encoding strategy is then decided to minimize both pre-roll delay and overall perceptual distortion of the transmitted video. This solution requires $K$ different coding results for each of the $N$ shots. For example, we can select 2 frame rates, 2 spatial resolutions and 3 quantization parameters in a typical application, which results in $K=12$. Then, the total storage requirement is the sum of the sizes of 12 compressed video streams. For a 90 minutes soccer video, where the average bitrate is 100 kbps over 12 encoded streams, the total storage is 791 MBytes. Although quantization parameter is embedded in the encoded bitstream, spatial resolution and frames per second may need to be sent as side information so as to synchronize when they are changed.

The operation of the decoder is straightforward. If the coding standard used supports spatio-temporal resolution changes, the resulting compressed bitstreams will be standards compliant. However, we may need a specialized display module to display all pictures at a standard spatial resolution. The display module may use the side information, consisting of the spatial and temporal resolution of each GoP to display the entire video using a single spatial and temporal resolution.

In the AVC/H.264 reference encoder, the HRD model assumes that the video will be drained by a CBR channel with a rate equal to the video encoding rate. Since in our proposed system, the target bitrates assigned to each segment varies, and for some segments the target encoding bitrate can be more than the channel rate, additional logical encoder buffer will be needed to store the excess bits produced. Because bits transmitted during the pre-roll time need to be stored at the decoder side, an identical additional logical buffer will be required at the decoder as well to ensure the proper operation of the proposed variable target rate system. The required additional logical buffers at the encoder and decoder are illustrated in Fig. 5. Here, the “logical” buffers demonstrate the necessary increase in the size of the codec buffers to realize DDO rate allocation. In an actual implementation, the “logical” buffers can be realized by simply increasing the codec buffer sizes accordingly.

**IV. EXPERIMENTAL RESULTS**

In our experiments, we used AVC/H.264 codec software JM 7.4 provided by the Joint Video Team (JVT) to encode each video segment using a number of fixed quantization parameters. We selected a 20 seconds soccer video clip, which is $352\times 288$ and 25 fps. The video is segmented into $N=4$ shots...
using the content analysis technique of [15]. The first shot is a goal event that is of great interest to most users, the second shot is a scene where the players cuddle to celebrate the goal. The audience is shown on the third shot and finally the team coach is seen on the last shot. We encoded each segment using spatial resolutions of 176×144 and 96×80, temporal resolutions of 25 fps, 12.5 fps and 6.25 fps and quantization parameters (QP) that vary between 17 and 36 for a total of \( K = 232 \) combinations. Here, \( K \) is chosen large to better study the trade-off along the convex Pareto-optimal curve on a fine scale. However, \( K \) can be reduced significantly by limiting the choice of quantization parameters to 2-3 values without noticeable performance degradation as shown in Fig. 11. We computed the total bits (rates) and distortion values (as a linear combination of the PSNR and blockiness measures given by Eqn.7) for each combination.

User-specified relevance values for the four shots used in our experiments and refined (final) weights scaled using audio information (audio energy distribution function given in Fig. 6) are shown in Table 1. Note that, in our formulation, ratios of weights (to each other) rather than the weights themselves are important. The relevance values can vary between different users. For example, if a user doesn’t want to see parts of video where only the audience is shown, the weight of that shot should be set to zero. In this case, the optimal encoding result may not include this irrelevant shot at all.

![Audio energy distribution of the whole video.](image)

**TABLE 1: WEIGHTS GIVEN BY THE USER AND REFINED (SCALED) BY AUDIO INFORMATION**

<table>
<thead>
<tr>
<th>Shot</th>
<th>Relevance</th>
<th>Average Audio Energy</th>
<th>Scaled Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.2689</td>
<td>0.759</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>0.2238</td>
<td>0.158</td>
</tr>
<tr>
<td>3</td>
<td>0.125</td>
<td>0.1582</td>
<td>0.056</td>
</tr>
<tr>
<td>4</td>
<td>0.125</td>
<td>0.0745</td>
<td>0.026</td>
</tr>
</tbody>
</table>

![Sample frames from each of the four shots that are adaptive DDO coded (on the left) and standard RDO coded (on the right) at 37.57 kbps.](image)

**TABLE 2: OPTIMAL SET OF PARAMETERS FOR THE VIDEO SEGMENTS**

<table>
<thead>
<tr>
<th>Shot</th>
<th>Scaled ( w_n )</th>
<th>Resolution</th>
<th>FPS</th>
<th>Bitrate</th>
<th>TD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.759</td>
<td>176×144</td>
<td>6.25</td>
<td>68.25 kbps</td>
<td>4.97 sec.</td>
</tr>
<tr>
<td>2</td>
<td>0.158</td>
<td>96×80</td>
<td>6.25</td>
<td>28.58 kbps</td>
<td>9.61 sec.</td>
</tr>
<tr>
<td>3</td>
<td>0.056</td>
<td>96×80</td>
<td>6.25</td>
<td>27.4 kbps</td>
<td>2.86 sec.</td>
</tr>
<tr>
<td>4</td>
<td>0.026</td>
<td>176×144</td>
<td>6.25</td>
<td>23.25 kbps</td>
<td>2.56 sec.</td>
</tr>
</tbody>
</table>

Fig. 7 shows a comparison of QCIF resolution key frames from different types of shots encoded by the proposed content adaptive DDO rate allocation technique (each GoP coded utilizing RDO) and the standard RDO codec (JM 7.4 from the JVT group) at the same rate. The encoding rate is 37.57 kbps, the channel rate and the available physical receiving buffer size are assumed to be 25 kbps and 50 kBytes, respectively; resulting in an average encoding rate of 37.57 kbps and an overall delay of 10.06 seconds for the content adaptive codec at the receiving side. While the ball and lines of the field are quite noticeable in the content adaptive DDO rate-allocated clip, we can’t see the ball and certain parts of the pitch lines in the standard RDO encoded version at 37.57 kbps. Also, for the 2nd shot, the blocking artifacts are very disturbing in the standard encoded version. For the last two shots, both coding schemes show similar performances. Fig. 8 and Fig. 9 show the quantization parameters and corresponding distortion measures, respectively, at each frame for both coding schemes.
Buffer Requirements: The proposed content-adaptive (DDO) results are compared with the variable bitrate (VBR) coded (using constant picture resolution and quantization factor for the whole video) and constant bitrate (CBR) coded (using AVC/H.264 rate control [9]) versions of 120 seconds long video obtained by cascading six identical replicas of the original video. We illustrate the instantaneous decoder buffer occupancies of DDO, and regular VBR and CBR solutions with equal average bitrate in Fig. 10, where we assure that the pre-roll delays are sufficient to guarantee continuous playback for each case for a fair comparison. The horizontal axis in Fig. 10 denotes the time elapsed after the encoder side starts streaming. For our content adaptive DDO solution, the changes in the buffer level can be either steep or slow depending on whether a high relevance or a low relevance segment is displayed. In our solution, the maximum physical client buffer size is set to 300 kBytes; although the maximum buffer level observed (necessary and sufficient client buffer size) is found to be 125 kBytes with a pre-roll delay of 40 seconds. The resulting average encoding bitrate throughout the video is 33.33 kbps. For this example, at the same bitrate, VBR and CBR solutions require approximately the same buffer size as our solution; hence, our solution provides higher video quality in important temporal segments without incurring additional pre-roll delay and buffer requirements over the standard CBR solution with RDO.

If the buffer constraint is kept too small, it may not always be possible to come up with a feasible solution, unless concessions are made on one or more of the constraints and/or objective functions. Note that the DDO solution can not be dominated by either CBR or VBR solutions both in pre-roll delay and distortion, since the optimal DDO solution would approach the better one of these solutions in a worst case scenario. In cases where there exist feasible solutions, our framework would find the optimal solution.

Delay-Distortion Trade-off: For a 25 kbps constant bitrate channel, the delay-distortion curves for the 120 seconds video with no buffer constraints imposed are shown in Fig. 11. For equal pre-roll delays, the proposed solution shows better weighted distortion performance on average, especially at the important temporal segments, for which the video PSNR gain is around 4.5 dB compared to the VBR solution. Note that, as the pre-roll delay increases, the required buffer size at the client size also has to increase. As a result, the larger the receiver buffer is, the more flexibility the encoder side has on GoP level bit allocation, increasing the overall video quality, as seen in Fig. 11.

In order to illustrate how the minimum required decoder buffer size is affected by the individual shot durations, we now construct a 1200 seconds (20 minutes) long video where the duration of each shot in the 120 seconds long video is made ten times larger (shot durations up to 96.1 seconds). For the same target bit rate (33.33 kbps), if the same bit allocation strategy among shots is applied as for the 120 seconds video, the required buffer size would be 1250 kBytes. On the other hand, if we re-run our optimization algorithm for the 1200 seconds video under the same conditions with a maximum physical buffer size constraint of 1000 kBytes., this results in an average video encoding bitrate of 30.03 kbps, a pre-roll delay of 243.9 seconds, and a minimum required buffer size of
762.2 kBytes. For the same average-bitrate video, the download-play approach would result in 1443.8 seconds of delay and 4512 kBytes of storage space. Hence, the pre-roll delay and buffer requirements of our method do not grow linearly with overall video or individual shot durations. On the contrary, the maximum physical buffer size constraint in the optimization formulation causes the average encoding bitrate to drop when necessary.

The results for 120 seconds and 1200 seconds videos, shown in Table 3 indicate that the required buffer sizes are well within the capabilities of today’s clients. With the continuous playback guarantee, the pre-roll delay and buffer size requirements of the standard RDO solution is very close to ours. However, the video quality is much higher in the important temporal segments in our solution. Hence, our solution provides higher video quality in important temporal segments without incurring additional pre-roll delay and buffer requirements (penalty) over the standard RDO solution.

<table>
<thead>
<tr>
<th>( R_b ) (kbps)</th>
<th>Avg. coding rate (kbps)</th>
<th>( TD ) (sec)</th>
<th>Download-play solution</th>
<th>Proposed Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Delay (sec)</td>
<td>Required Buffer (kB)</td>
</tr>
<tr>
<td>25</td>
<td>33.33</td>
<td>120</td>
<td>160</td>
<td>500</td>
</tr>
<tr>
<td>25</td>
<td>30.08</td>
<td>1200</td>
<td>1443.8</td>
<td>4512</td>
</tr>
</tbody>
</table>

The results presented here are provided as a proof of concept. Improvements in weighted PSNR and pre-roll delay may vary with the video content and specific relevance-distortion policy adopted.

V. CONCLUSIONS

This paper introduces a new MOO framework for delay-distortion optimization (DDO) in content-based adaptive GoP-level rate allocation for video streaming over resource-limited networks using linear programming. Semantic relevancy of shots has been taken into account in determination of encoding parameters for each shot. Clearly, video with unacceptable quality is by definition of no use for anyone. On the other hand, there are users who will wait to watch video at an acceptable quality as manifested by the streaming applications on the Internet. What we accomplished in this paper is that, we developed a technique to reduce this waiting time to levels much lower than that of download and play, keeping the relevant quality at an acceptable level over low bandwidth channels. The proposed method not only maximizes perceptual quality of relevant parts in the video, but also minimizes the pre-roll (initial playback) delay at the receiving side. It outperforms the performance of regular bit allocation schemes in the relevant shots (4.5 dB gain), while still providing an acceptable quality for other shots with quite affordable buffer requirements. The proposed framework does not depend on a particular video coding technology.

APPENDIX

MULTIPLE-OBJECTIVE OPTIMIZATION

A thorough treatment of multiple-objective optimization (MOO) techniques can be found in [29]-[30]. Here, we present a simple example with two objectives to demonstrate the main idea of MOO, where we

\[
\min_{x,y} f(x,y) = \min_{x,y} x \cdot y \\
\min_{x,y} g(x,y) = \min_{x,y} \frac{200}{x} + \frac{200}{y}
\]

jointly subject to \( x \in [1,20] \) and \( y \in [1,20] \).

The sketch of the functions \( f(x,y) \) and \( g(x,y) \) for the region of interest is shown in Fig. 12. The point \((x,y) = (1,1)\) minimizes \( f \) with a minimum value of \( f_{\text{min}}=1 \) while \( g \) attains its maximum value, \( g_{\text{max}}=400 \) at this point. The other endpoint \((x,y) = (20,20)\) minimizes \( g \) with a minimum value of \( g_{\text{min}}=20 \), while \( f \) attains its maximum value \( f_{\text{max}}=400 \) at this point. A solution is called Pareto-optimal if any one of the objective values cannot be improved without degrading other objective values. In other words, a Pareto-optimal solution cannot be dominated (outperformed in all the objective functions) by any other feasible solution. In order to draw the Pareto-optimality trade-off curve shown in Fig. 13.

![Fig. 12. Sketch of the two functions f and g in the region of interest.](image)

![Fig. 13. Minimum values that the cost function g can take for possible values of f in the interval [f_{\text{min}}, f_{\text{max}}].](image)

An infeasible point that minimizes both of the objective functions individually, the point \((f_{\text{min}}=1, g_{\text{min}}=20)\) for our example, is called the utopia point. The best compromise solution is defined as the point on this curve that is closest to the utopia point \((f=1,g=20)\) in the Euclidian-distance sense after proper scaling (subtracting the mean and dividing by the standard deviation) of all feasible points. In our example, the closest point to the utopia point on this curve can be found as \((f=38.21, g=64.71)\). The corresponding \( x \) and \( y \) values are...
5-STRM Multimedia Streaming

determined as \( x = y = 6.181 \).

REFERENCES


