END-TO-END RATE-DISTORTION OPTIMIZED MODE SELECTION FOR MULTIPLE DESCRIPTION VIDEO CODING

Brian A. Heng[†]*, John G. Apostolopoulos*[◊]*, and Jae S. Lim*[†]

[†] Massachusetts Institute of Technology
Cambridge, MA, USA

[◊] Streaming Media Systems Group
Hewlett-Packard Labs, Palo Alto, CA, USA

ABSTRACT

Multiple description (MD) video coding can be used to reduce the detrimental effects caused by transmission over lossy packet networks. Each approach to MD coding consists of a tradeoff between compression efficiency and error resilience. How effectively each method achieves this tradeoff depends on the network conditions as well as on the characteristics of the video itself. This paper proposes an adaptive MD coding approach which adjusts to these conditions through the use of adaptive MD mode selection. The encoder in this system is able to accurately estimate the expected end-to-end distortion, accounting for both coding and packet-loss-induced distortions, as well as for the bursty nature of channel losses and the effective use of multiple transmission paths. With this model of end-to-end expected distortion, the encoder selects between MD coding modes in a rate-distortion optimized manner to most effectively trade-off compression efficiency for error resilience. We show how this approach adapts to the local characteristics of the video as well as to current network conditions and demonstrate the resulting gains in performance.

1. INTRODUCTION

Streaming video applications often require error resilient coding methods able to adapt to current network conditions and to withstand transmission losses. Best-effort networks like the Internet are characterized by variable bandwidths, packet losses, and delays. Applications must be able to withstand these harsh conditions or they can suffer severe performance degradations.

Multiple description (MD) video coding is one approach that can be used to reduce the detrimental effects caused by packet loss on best-effort networks. In a multiple description system, a video sequence is coded into two or more complementary streams in such a way that each stream is independently decodable. The quality of the received video improves with each received description, but the loss of any one of these descriptions does not cause complete failure. If a portion of one of the streams is lost or delivered late, the video playback can continue with only a slight reduction in overall quality. For an in-depth review of MD coding for video communications see [1].

Previous MD video coding approaches applied a single MD technique to an entire sequence. However, the optimal MD coding method depends on many factors including the amount of motion in the scene, the amount of spatial detail, desired bitrates, error recovery capability of each technique, current network conditions, etc. This paper examines the adaptive use of multiple MD coding modes within a single sequence. Specifically, this paper proposes an adaptive MD coder which selects among MD coding modes in an end-to-end rate-distortion (R-D) optimized manner as a function of local video characteristics and network conditions. The addition of end-to-end R-D optimization is an extension of the adaptive system proposed in [2].

This paper continues in Sections 2 and 3 with an overview of how end-to-end optimized mode selection can be achieved in MD systems. The details of the proposed system are provided in Section 4, and experimental results are given in Section 5.

2. OPTIMAL MD MODE SELECTION

Each approach to MD coding trades off some amount of compression efficiency for an increase in error resilience. How efficiently each method achieves this tradeoff depends on the quality of video desired, the current network conditions, and the characteristics of the video itself. Most prior work in MD coding apply a single MD method to the entire sequence; this approach is taken so as to evaluate the performance of each MD method. However, it would be more efficient to adaptively select the best MD method based on the situation at hand. Since the encoder in this system has access to the original source, it is possible to calculate the rate-distortion statistics for each coding mode and select between them in an R-D optimized manner.

Lagrangian optimization techniques can be used to minimize distortion subject to a bitrate constraint [3]. However, this approach assumes the encoder has full knowledge of the end-to-end distortion experienced by the decoder. In a lossy channel, the end-to-end distortion consists of (1) known distortion from quantization and (2) unknown distortion from random packet loss which can only be determined in expectation due to the random nature of losses. Modifying the Lagrangian cost function to account for the total end-to-end distortion gives the following.

$$J(\lambda) = D_i^{quant} + E\left[\tilde{D}_i^{loss}\right] + \lambda R_i \qquad (1)$$

Here R_i is the total number of bits necessary to code region i, D_i^{quant} is the distortion due to quantization, and $\tilde{D}_i^{loss}$ is a random variable representing the distortion due to packet losses. Thus, the expected distortion experienced by the decoder can be minimized by coding each region with all available modes and choosing the mode which minimizes the Lagrangian cost.

Calculating the expected end-to-end distortion is not a straightforward task due to spatial and temporal error propagation. However, in [4] the authors show how to estimate expected distortion in a pixel-accurate recursive manner for SD and Bernoulli losses. In the next section we discuss this approach and the extensions necessary to apply it to the current problem of MD coding over multiple paths with Gilbert (bursty) losses.

3. MODELING EXPECTED DISTORTION IN MULTIPLE DESCRIPTION STREAMS

As discussed in Section 2, random packet losses force the encoder to model the network channel and estimate expected end-to-end distortion. With an accurate model of expected distortion the encoder can make optimized decisions to improve the quality of the decoded video stream. A number of approaches have been suggested in the past to model expected distortion. In [4] the authors suggest a recursive optimal per-pixel estimate (ROPE) for optimal intra/inter mode selection. In [5] the ROPE model is extended to a two-stream multiple description system by recognizing the four possible loss scenarios: both descriptions are received, either description is lost, or both descriptions are lost. The conditional expectations of each of these four possible results are multiplied by the probability of each occurring to calculate the total expectation.

Previous models have used a Bernoulli independent packet loss model, but the idea can be modified for a channel with bursty packet losses as well. Recent work has identified the importance of burst length in characterizing error resilience schemes. In fact burst length has been shown to be an important feature for comparing the relative merits of different error resilient coding schemes [6][7][8].

For this system we have extended the MD ROPE approach to account for bursty packet loss. Here we use a 2-state Gilbert loss model, but the same approach could be used for any multi-state model including those with fixed burst lengths. In the Gilbert loss model, packet losses become more likely if the previous packet has been lost. The total expectation can be calculated with multi-state packet loss models by computing the expectation conditioned on being in each state and multiplying by the probability of transitioning from one state to another. We have further modified this approach in order to apply it to H.264 with quarter pixel motion vector accuracy and more sophisticated error concealment methods by using the techniques proposed in [9] for estimation of cross-correlation terms.

4. SYSTEM IMPLEMENTATION

The system described in this paper has been implemented based on the H.264 video coding standard using quarter pixel motion vector accuracy and all available intra- and inter-prediction modes [10]. We have used reference software version 8.6 for these experiments with modifications to support adaptive mode selection. Constant bitrate encoding is used to keep the number of bits per frame approximately constant. To accomplish this, the quantizer for each macroblock is adjusted using the reference software rate-control implementation. The in-loop deblocking filter used in H.264 has been turned off to simplify the problem.

The adaptive mode selection is performed on a macroblock basis using the Lagrangian techniques discussed in Section 2 with the expected distortion model from Section 3. Note that this optimization is performed simultaneously for both traditional coding decisions (e.g. inter versus intra coding) as well as for selecting one of the possible MD modes.

The current system uses a combination of four possible MD modes: single description coding (SD), temporal splitting (TS), spatial splitting (SS), and repetition coding (RC). SD coding represents the typical coding approach where frames are predicted from the previous frame in an attempt to remove as much redundancy as possible. In temporal splitting mode, even frames are predicted from even frames and odd frames from odd frames. Similarly, in spatial splitting, even lines are predicted from even lines and odd from odd (it was necessary to modify the H.264 codec to support macroblock-level interlaced coding for this approach). Finally, repetition coding is the same as the SD approach except the data is transmitted once in each description.

Note that when coded in a non-adaptive fashion, each method (SD, TS, SS, RC) is still performed in an R-D optimized manner as mentioned above. All of the remaining coding decisions, including inter versus intra coding, are made to minimize the end-to-end distortion. For instance, the RC mode is not simply a straightforward replica of the SD mode. The system recognizes the reliability of the RC mode and elects to use far less intra-coding allowing more intelligent allocation of the available bits.

The packetization of data differs slightly for each mode (see Fig 1). In both the SD or TS approaches, all data for a frame is placed into a single packet. The even frames are then sent along one stream and the odd frames along the other. While in the SS and RC approaches, data from a single frame is coded into packets placed into both streams. Even lines are sent in one stream and odd lines sent in the other with SS, while all data is repeated for RC. Therefore, for SD and TS each frame is coded into one large packet which is sent in alternating streams, while for SS and RC each frame is coded into two smaller packets and one small packet is sent in each stream. Since the adaptive approach (ADAPT) is some combination of each of these four methods, there is typically one slightly larger packet and one smaller packet and these alternate streams between frames.

If a frame is lost in either the TS or SD method, no data exists in the opposite stream at the same time instant, so missing data is directly copied from the previous frame. In the SS method, if only one description is lost the decoder reconstructs missing lines using linear interpolation, and if both are lost it copies the previous frame. Similarly for RC, if only one description is lost the decoder can use the data in the opposite stream, while if both are lost it copies the previous frame.

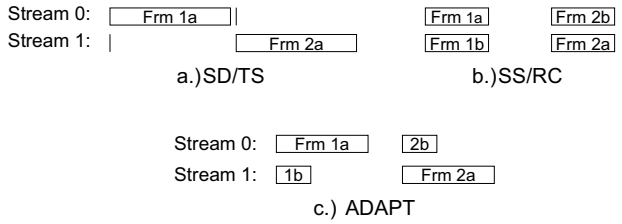

Fig 1: Packetization of data in MD modes. a.) SD/TS: Data sent along one path alternating between frames. b.) SS/RC: Data spread across both streams. c.) ADAPT: Combination of the two resulting in one slightly larger packet and one slightly smaller.

5. EXPERIMENTAL RESULTS

These results have been obtained using our modified H.264 JM 8.6 codec (described above) and the Foreman video sequence, which contains 400 frames at 30 frames per second at QCIF resolution.

To measure the actual distortion experienced at the decoder, we have simulated a Gilbert packet loss model with packet loss rates and expected burst lengths as specified in each section

below. For each of the experiments, we have run the simulation with 300 different packet loss traces and averaged the resulting squared-error distortion. The same packet loss traces were used throughout a single experiment to allow for meaningful comparisons across the different MD coding methods.

Each path in the system is assumed to carry 30 packets per second where the packet losses on each path are modeled as a Gilbert process. For wired networks, the probability of packet loss is generally independent of packet size so the variation in sizes should not generally affect the results or the fairness of this comparison. When the two paths are balanced or symmetric the optimization automatically sends half the total bitrate across each path. For unbalanced paths the adaptive system results in a slight redistribution of bandwidth as discussed below.

We first evaluate the system's ability to adapt to the characteristics of the video source. The channel in this experiment was simulated with two balanced paths each having 5% average packet loss rate and expected burst length of 3 packets. The video was coded at approximately 0.4 bits per pixel (bpp). Fig 2 demonstrates the resulting distortion in each frame averaged over the 300 packet loss traces for the adaptive MD method and each of its non-adaptive MD counterparts.

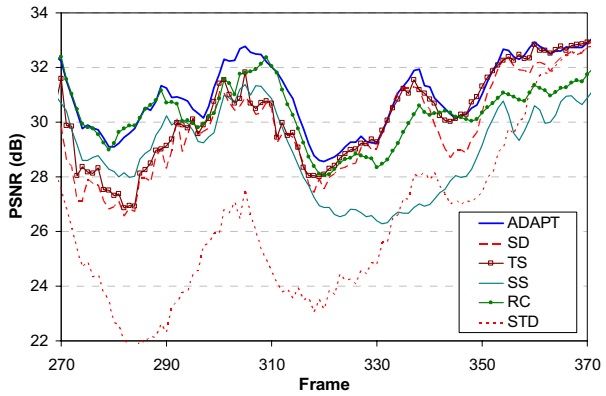

Fig 2: Average distortion in each frame for ADAPT versus each non-adaptive approach. Coded at 0.4 bpp with balanced paths and 5% average packet loss rate and expected burst length of 3.

The Foreman sequence contains a significant amount of motion from frames 250 to 350 and is fairly stationary from frame 350 to 399. Notice how the SS/RC methods work better during periods of significant motion while the SD/TS methods work better as the video becomes stationary. The adaptive method intelligently switches between the two, maintaining at least the best performance of any non-adaptive approach. Since the adaptive approach adapts on a macroblock level, it is often able to do even better than the best non-adaptive case by selecting different MD modes within a frame as well.

Also shown on Fig 2 are the results from a typical video coding approach which we will refer to as standard video coding (STD). Here R-D optimization is only performed with respect to quantization distortion, not the end-to-end R-D optimization used in the other approaches. Instead of making inter/intra coding decisions in an end-to-end R-D optimized manner as performed by SD, it periodically intra updates one line of macroblocks in every other frame to combat error propagation (this update rate was chosen as the optimal intra refresh rate [11] is often approximately $1/p$, where p is the packet loss rate).

By making intelligent decisions through end-to-end R-D optimization, the SD method is able to outperform the STD method by as much as 4 or 5 dB. The adaptive MD approach is further able to outperform SD coding by up to 2 dB depending on the amount of motion present at the time.

Fig 3 shows the percentage of macroblocks using a particular MD mode in each frame. From the distribution of MD modes, one can roughly segment the sequence into three distinct regions: almost exclusively SD/TS in the last 50 frames, mostly SS/RC in the middle, and a combination of the two at the beginning. This matches up with the characteristics of the video which contains some amount of motion at the beginning, a fast camera scan in the middle, and is nearly stationary at the end.

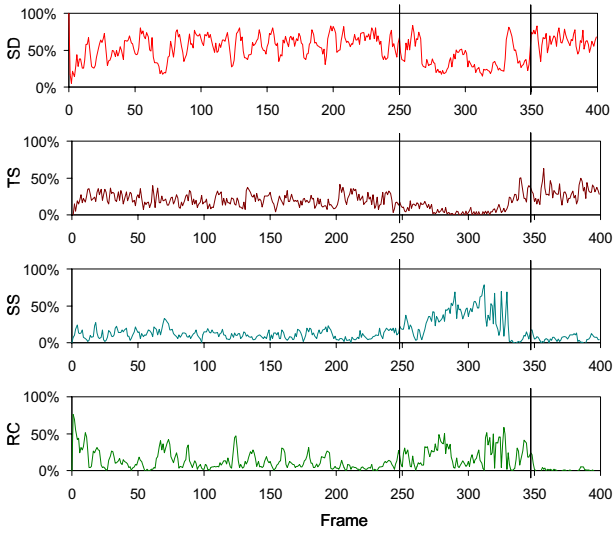

Fig 3: Distribution of MD modes used in adaptive method for each frame. 5% average packet loss rate, expected burst length 3.

In our second experiment we examine how the system adapts to the conditions of the network. Here we have compared the previous experiment with one in which the average packet loss rate is increased to 10%. Table 1 shows the distribution of MD mode in each of these cases. As the loss rate increases to 10% the system responds by switching from lower redundancy methods (SD/TS) to higher redundancy methods (SS/RC) in an attempt to provide more protection against losses.

Fig 4 shows the end-to-end R-D performance curves of each method. To generate each point on these curves, the resulting distortion was averaged across all 300 packet loss simulations, as well as across all 400 frames of the sequence. The same calculation was then conducted at various bitrates to generate each R-D curve. By switching between MD methods, ADAPT is able to outperform optimized SD coding by 0.2-1.2 dB and STD coding by as much as 4.9 dB. ADAPT is able to outperform TS, which performs second best overall, by as much as 0.6 dB.

Table 1: Distribution of MD modes in the adaptive approach comparing 5% and 10% average packet loss rates.

MD Mode	Low Loss	High Loss
SD	51.9%	43.9%
TS	19.4%	17.5%
SS	15.4%	16.8%
RC	13.4%	21.9%

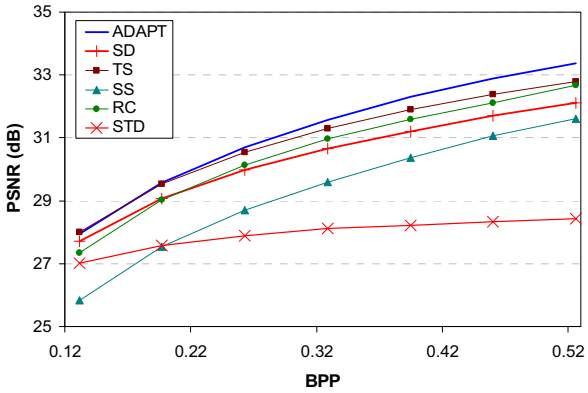

Fig 4: End-to-end R-D performance of ADAPT and non-adaptive methods. 5% packet loss rate, expected burst length 3.

Table 2: Percentage of macroblocks using each MD mode in the adaptive approach when sent along unbalanced paths.

MD Mode	Even Frames More Reliable Path	Odd Frames Less Reliable Path
SD	55.6%	48.8%
TS	25.2%	14.5%
SS	10.7%	18.6%
RC	8.4%	18.1%

Table 3: Percentage of total bandwidth in each stream for balanced and unbalanced paths.

	Balanced Paths	Unbalanced Paths
Stream 1	49.9%	56.2%
Stream 2	50.1%	43.8%

One interesting side result here is how well RC performs at higher bitrates. Keep in mind that this is an R-D optimized RC approach, not simply the half-bitrate SD method repeated twice. The amount of intra coding used in RC is significantly reduced relative to SD coding as the encoder recognizes the increased resilience of the RC method and chooses to allocate more bits for improving quality.

In our final experiment, we analyze the performance of the adaptive method when used with unbalanced paths where one path is more reliable than the other. The channel consisted of one path with 3% average packet loss rate and another with 7%, both with an expected burst length of 3 packets. The video in this experiment was coded at approximately 0.4 bpp. Table 2 shows the distribution of MD modes in even frames of the sequence versus odd frames. The even frames are those where the larger packet (see Fig 1) is sent along the more reliable path and the smaller packet is sent along the less reliable path. The opposite is true for the odd frames.

As shown in Table 2, the system uses more SS and RC in the less reliable odd frames. These more redundant methods allow the system to provide additional protection for those frames which are more likely to be lost. By doing so, the adaptive system is effectively moving data from the less reliable path into the more reliable path. Table 3 shows the bit rate sent along each path in the balanced versus unbalanced case. In this situation, the system is shifting about 6% of its total rate into the more reliable stream to compensate for conditions on the network. Since the non-adaptive methods are forced to send approximately half their total rate along each path, it is difficult to make a fair comparison across methods in this unbalanced situation. We are considering ways to compensate for this. However, it is quite interesting that the end-to-end R-D optimization is able to adjust to this situation in such a manner.

6. CONCLUSIONS

This paper proposed an end-to-end R-D optimized adaptive mode selection system for multiple description coding. The system makes use of multiple MD coding modes within a given sequence, making optimal decisions using a model of expected end-to-end distortion. We have demonstrated how the system is able to adapt to local characteristics of the video and to network conditions on multiple paths and have shown the potential for this adaptive approach, which selects among a small number of simple complementary MD modes, to significantly improve video quality. The effectiveness of this adaptive scheme depends on the video source and knowledge of the network. Even so, the results are quite promising, and it is apparent that the adaptive MD mode selection can provide significant benefits.

7. REFERENCES

[1] Y. Wang, A. Reibman, and S. Lin, "Multiple description coding for video communications," to appear in *Proceedings of the IEEE*, 2005.

[2] B. Heng and J. Lim, "Multiple description video coding through adaptive segmentation," *Proc. of SPIE*, vol. 5558, Aug. 2004.

[3] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, Nov 1998.

[4] R. Zhang, S. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, June 2000.

[5] A. Reibman, "Optimizing multiple description video coders in a packet loss environment," in *Packet Video Workshop*, April 2002.

[6] J. Apostolopoulos, "Reliable video communication over lossy packet networks using multiple state encoding and path diversity," in *Proc. SPIE, VCIP*, pp. 392-409, January 2001.

[7] J. Apostolopoulos, W. Tan, S. Wee, and G. Wornell, "Modeling path diversity for multiple description video communication," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, May 2002.

[8] Y. Liang, J. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: Does burst-length matter?," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 5, pp 684-687, 2003.

[9] H. Yang and K. Rose, "Recursive end-to-end distortion estimation with model-based cross-correlation approximation," in *Proc. of the IEEE Int. Conf. on Image Processing*, vol. 3, pp. 469-472, September 2003.

[10] ITU-T Rec. H.264, "Advanced video coding for generic audiovisual services", March 2003.

[11] K. Stuhlmüller, N. Färber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1012-1032, June 2000.