Submitted to Int. Conf. Image Proc., Barcelona, Spain, 2003

PRESCIENT R-D OPTIMIZED PACKET DEPENDENCY MANAGEMENT FOR LOW-LATENCY VIDEO STREAMING

Yi J. Liang and Bernd Girod

Information Systems Laboratory, Department of Electrical Engineering
Stanford University, Stanford, CA 94305-9510, USA
{yiliang, bgirod}@stanford.edu

ABSTRACT

We present an improved, prescient, scheme to deliver pre-encoded video streams at very low latency. We dynamically manage the prediction dependency for a group of packets using a rate-distortion (R-D) optimization framework that adapts to the channel, so that the expected end-to-end distortion is minimized under a given rate constraint. In our system, the R-D characteristics of each prediction mode are pre-computed at compression time and stored in an R-D preamble along with the multiple compressed versions of the stream. The distortion in the preamble is estimated using an accurate loss-distortion model, and the optimal prediction modes are computed with an iterative descent algorithm. Complexity is reduced at streaming time by using the preamble to compute the optimal modes instead of actually compressing the video. Experiments demonstrate that the proposed prescient scheme provides significant performance gains over simple Intra-insertion, and consistent bitrate savings compared to the greedy scheme.

1. INTRODUCTION

Internet video streaming today is plagued by variability in throughput, packet loss, and delay due to network congestion and the heterogeneous infrastructure. To mitigate these effects, media streaming systems typically employ a large receiver buffer that introduces a latency of 10-15 seconds. This is undesirable since the slow start-up is annoying and high latency severely impairs the interactive playback features, such as VCR functionality.

Low latency in video streaming is difficult due to the lossy nature of the transmission channel in combination with the sensitivity of the compressed video stream against packet losses. For a typical hybrid video codec, an Inter-coded frame is predicted from a reference picture with motion compensation, so that the temporal redundancy across successive pictures is removed or reduced to provide higher coding efficiency. However, proper decoding of such Inter-coded pictures depends on the error-free reception and reconstruction of the reference picture it uses, which is not guaranteed over lossy channels.

Assume the typical scenario where an IP packet contains one video frame. If a packet (frame) is lost, the proper reconstruction of all subsequent frames that depend on the lost frame is affected. Hence, in a typical ARQ-based system, whenever a packet is lost, retransmission is required to guarantee the correct reception of each frame. The time for retransmissions constitutes the major part of the total end-to-end delay. Other error-resilient schemes recently proposed include packet transmission scheduling [1][2], forward error correction (FEC) [3], and long-term memory (LTM) prediction [4]. These schemes introduce lower latency than ARQ, but at various delay and rate-distortion (R-D) costs. All the schemes above provide improved error-resilience through source/channel coding and optimizing the transmission policy, which may include transmission/retransmission schedules, the amount of redundancy in an FEC system, and the prediction dependency across packets.

In this work, we consider applications that impose very stringent delay requirements. In such scenarios, packets are transmitted as soon as they become ready at the sender, and retransmissions cannot be used. In order to increase error-resilience and eliminate the need for retransmission, we pre-store multiple representations of certain frames at the server so that a representation can be chosen that only uses reference frames that will be available with very high probability. We account for the dependency across packets resulting from hybrid video coding and dynamically manage this dependency to achieve increased error-resilience given the data rate constraint. Once the need for retransmission is eliminated, buffering is needed only to absorb the packet delay jitter, so that the streaming latency can be reduced to a few hundred milliseconds.

In our previous work [5], we optimize the prediction dependency of the *next* packet to transmit for live-encoding. The optimality achieved with this *greedy* method is local. In this work we optimize the prediction dependencies for a *group* of packets before they are transmitted, by using the pre-computed R-D information for pre-compressed streams. This approach, which we refer to as the *prescient* method, may achieve global R-D optimality for a group of packets, or even the entire sequence and it can be realized at low complexity in practical systems. A prescient scheme for selecting the optimal Intra/Inter coding mode for macroblocks was ever proposed in [6], which achieves optimality for current and a limited number (one or two) of subsequent frames. In this work, we consider a larger group of more frames in determining the optimal prediction dependency. The estimation of distortion based on different loss events, as well as determining the prediction modes with a low-complexity algorithm are the major contributions of this work.

This paper is structured as follows: we first describe dynamic management of packet dependencies. In Section 3, we present prescient packet dependency management using the R-D preamble, and the simplified algorithms for rate and distortion estimation and determining the prediction modes. Experimental results are presented in Section 4.

This work has been supported by a gift from HP Labs, Palo Alto, CA.

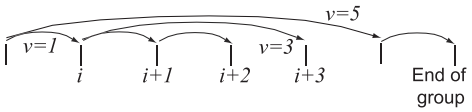

Fig. 1. Illustration of prediction modes.

2. DYNAMIC MANAGEMENT OF PACKET DEPENDENCIES FOR VIDEO STREAMING

In a conventional coding and transmission scheme, an I-frame is typically followed by a series of P-frames, which are predicted from their immediate predecessors. This fixed coding structure is vulnerable to channel errors since each P-frame depends on its predecessors, and any packet loss will break the prediction chain and affect all subsequent P-frames. If, instead, a frame is predicted from a more reliable frame that precedes the most recent frame, the scheme is more robust against channel errors. Let $v = 1, 2, 3, ...V$ denote the *prediction mode* (or *coding mode*) chosen for a video frame, where v indicates which previous frame is chosen for prediction, as illustrated in Fig. 1. V is the number of previous decoded frames available from the LTM, and $v = 0$ denotes Intra-coding. Using a prediction mode $v > 1$ increases error-resilience [5]. The robustness is normally obtained at the expense of a higher bitrate since the correlation between two frames in general becomes weaker as they are more widely separated.

In this work, we select the reference at the frame level and assume that each predictively coded frame is coded into one IP packet (the proposed scheme can also be extended to the case where a frame is coded into multiple packets). In this way we manage the frame prediction dependency at the packet level. We consider the dependency across packets resulting from hybrid video coding and dynamically manage this dependency while adapting to the varying channel conditions. Due to the trade-off between error-resilience and coding efficiency, we apply *Optimal Reference Picture Selection (ORPS)* within an R-D framework, by considering video content, channel loss probability and channel feedback (e.g., ACK, NACK, or time-out) [5]. When coding a frame, several versions are generated, including Intra-coding as well as Inter-coding using different reference modes, e.g., $v = 1, 2, 3, ...$. The associated rate and expected distortion are obtained and stored as auxiliary information, referred to as the *R-D preamble*, for later use in determining the optimal coding modes that minimize the R-D costs. The proposed scheme is compatible with the emerging ITU-T H.264 standard [7].

3. PRESCIENT PACKET DEPENDENCY MANAGEMENT USING R-D PREAMBLES

3.1. Prescient packet dependency management

For streaming of pre-compressed video, we take advantage of pre-computed rate and distortion information in the R-D preamble, which is actually a compact description of packet contents. With the auxiliary information computed and stored offline, we are able to determine the R-D optimized coding modes by only considering the R-D preamble instead of actually compressing the video, which greatly reduces the complexity at streaming time.

Unlike the greedy method in our prior work [5], we predetermine the coding modes for a *group* of frames before they are sent. This group of frames can be a GOP starting with a traditional I-frame, or an instantaneous decoder refresh (IDR) picture proposed in H.264 [7], which resets the multiframe buffer to break the inter-dependencies from any picture decoded prior to the IDR picture. In this case the group is independent of previous groups sent if its leading frame is properly decoded. The group of frames can also be a sub-GOP as we proposed in [8], and the leading frame of the group is predicted from the leading frame of the last sub-GOP. In this case, all the frames except the leading frame of the group are independent of previous groups.

For a group of L frames indexed by $i = 0, 1, ...L-1$, we use $\mathbf{v} = (v_0, v_1, ...v_{L-1})$ to denote their prediction modes. For the leading frame of the group, $i = 0$, we restrict its coding mode to be either Intra or $v_0 = L$, and the mode to use is determined by examining respective R-D costs, using any feedback information available. The i-th frame following the leading frame is allowed to use coding modes $1 \leq v_i \leq i$ and Intra, e.g. is only allowed to be predicted from prior frames in the same group. This keeps different groups relatively independent of each other, and helps to avoid potential mismatch error that might occur when assembling and transmitting pre-encoded bitstreams [8]. The coding modes of the frames following the leading frame are determined in a prescient fashion. We use $R_{\mathbf{v}}$ to denote the average rate for coding the group using modes $\mathbf{v}$, and $\overline{D}_{\mathbf{v}}$ to denote the corresponding expected distortion, given the channel conditions and the availability of packets for decoding. The Lagrangian cost associated with using a particular set of coding modes, $\mathbf{v}$, is given by

$$J_{\mathbf{v}} = \overline{D}_{\mathbf{v}} + \lambda R_{\mathbf{v}}. \quad (1)$$

The optimal prediction modes for the group is determined by searching for the combination that results in minimal $J_{\mathbf{v}}$

$$\mathbf{v}_{opt} = \arg\min_{\mathbf{v}} J_{\mathbf{v}}. \quad (2)$$

In (1), λ is a Lagrange multiplier. We use $\lambda = 5e^{0.1Q}(\frac{5+Q}{34-Q})$, which is the same as λ_{mode} in H.26L TML 8 used to select the optimal prediction mode [9], and Q is the quantization parameter used to trade off rate and distortion. At streaming time, when a prior frame that affects future dependency structure is negatively acknowledged by feedback, or time-out, the prediction dependency of the next frame to send is immediately changed by re-encoding using a most recently acknowledged frame as reference, or Intra-coding. Error propagation is stopped in this way, and the prediction modes for the rest of the frames in the group are re-computed to avoid dependencies on any prior loss-afflicted frames.

Compared with the greedy approach, the prescient scheme may achieve global R-D optimality for a group of frames (which can be extended to the entire sequence). Within a particular group, as earlier frames are more important in the dependency structure than later frames, the prescient scheme applies stronger error-protection (by using higher modes) on earlier frames, compared to the equal protection given by the greedy scheme. Due to the weaker protection applied to later frames in a group, at lower costs, we expect bitrate savings by using the prescient approach.

3.2. Rate and distortion estimation for generating R-D preambles

In generating the R-D preambles and determining the optimal prediction modes using (1), accurate estimation of the rates and distortions is critical. The rates corresponding to different coding modes

are obtained by encoding the frames with multiple trials offline. Denoting the rate of coding a particular frame i using mode v by $R_v[i]$, the average rate of the frames in a group with coding modes $\mathbf{v} = (v_0, v_1, ...v_{L-1})$ is given by

$$R_{\mathbf{v}} = \frac{1}{L}\sum_{i=0}^{L-1} R_{v_i}[i]. \quad (3)$$

Estimation of the distortion is more challenging since the reconstruction distortion depends on the loss events of the packets. For a group of L packets, there exist 2^L loss patterns, and it is not feasible to measure and tabulate the distortion values for all combinations of packet losses. We estimate the distortion by using an accurate loss-distortion model we proposed in our recent work [10]. With this model, distortion values for general loss patterns can be extrapolated from a limited number of measurements.

This model explicitly considers the effect of different loss patterns, including burst losses and separated losses spaced apart by a lag, and accounts for the correlation between error frames. A simple loss concealment scheme is assumed where the lost frame is replaced by the previous frame at the decoder output. It is found in [10] that the distortion produced by a burst loss is generally greater than the sum of an equal number of single isolated losses, since it also includes cross-correlation terms between the error frames.

To avoid computing the distortion of all 2^L loss patterns, we approximate the overall expected distortion by only considering the distortion of single losses that are independent of each other, as well as the cross-term between any *two* losses. We ignore the distortion resulting from the interaction between more than two losses, since the probability of having more than two losses in a group is much smaller. In [3], it is shown that an approximation using 1st and 2nd order Taylor expansion gives accurate estimation of the end-to-end distortion for packet loss rates of up to 20%. In this work, we base the distortion estimation on the model we established in [10]. For a particular group of L frames, the average mean-square distortion per frame is given by

$$\begin{aligned}\overline{D}_{\mathbf{v}} \quad &\approx \quad \overline{D}_Q + \frac{1}{L}(\sum_{i=0}^{L-1} D_{\mathbf{v}}[i]\cdot p[i] + \\ &\sum_{i=0}^{L-2}\sum_{j=i+1}^{L-1} \Delta D_{\mathbf{v}}[i][j]\cdot p[i]\cdot p[j]), \quad (4)\end{aligned}$$

where $\overline{D}_Q$ is the average distortion of the quantization error, and $p[i]$ is the loss probability of packet i. $D_{\mathbf{v}}[i]$ is the *total distortion* (not including the quantization error) of all the frames in the group as a result of losing packet i, which includes not only the mean-square error (MSE) of Frame i, but also error propagation within the group. Due to the manipulation of packet dependency, $D_{\mathbf{v}}[i]$ is a function of the prediction modes $\mathbf{v}$. $D_{\mathbf{v}}[i]$ is computed by

$$D_{\mathbf{v}}[i] = \sigma^2[i]\cdot\alpha_{\mathbf{v}}[i], \quad (5)$$

where $\sigma^2[i]$ is the MSE of Frame i, and $\alpha_{\mathbf{v}}[i]$ is a factor that takes the error propagation into account. In the example illustrated in Fig. 1, $\alpha_{\mathbf{v}}[i] = 1 + r + r^2 + r$, where r ($r < 1$) is a factor that accounts for the effect of spatial filtering in reducing the error power initially introduced. r mainly depends on the strength of the loop filter of the codec and is approximated by a constant average for a group or even for the entire sequence. In this example, the error introduced at Frame i is only propagated to three subsequent frames due to the prediction modes used.

In (4) $\Delta D_{\mathbf{v}}[i][j]$ is the cross-term in the total distortion as a result of losing both Frames i and j, which is not included in the distortions of single losses. $\Delta D_{\mathbf{v}}[i][j]$ is given by

$$\Delta D_{\mathbf{v}}[i][j] = 2\rho_{i,j}\cdot\sigma[i]\cdot\sigma[j]\cdot\alpha_{\mathbf{v}}[j], \quad (6)$$

where $\rho_{i,j}$ is the correlation coefficient between possible error Frames $j-1$ (propagated from Frame i), and j, more details of which can be found in [10]. If Frame $j-1$ is not afflicted by any error propagated from i, $\rho_{i,j} = 0$, due to the lack of any interaction between the two losses and the previous-frame-based concealment scheme.

To summarize, the three terms on the right-hand-side of (4) represent the quantization error free of frame losses, the distortion of single and independent losses, and the cross-term in the distortion of two losses, respectively. For a group of L frames, this approximation reduces the distortion computation from 2^L loss events down to $L + \frac{1}{2}L(L-1)$ events. The auxiliary information that needs to be pre-measured and stored includes:

(i) L MSE values $\sigma^2[i]$ for single losses at $i = 0, 1, ..., L-1$;

(ii) $\frac{1}{2}L(L-1)$ correlation coefficients $\rho_{i,j}$;

(iii) One attenuation factor r. The parameters in (i) - (iii) are to be used in (5), (6) and (4) to obtain $\overline{D}_{\mathbf{v}}$;

(iv) $\frac{1}{2}L(L+1)+1$ rate values, $R_v[i]$, for all eligible prediction modes, which are used in (3) to obtain $R_{\mathbf{v}}$.

3.3. Algorithm

With the coding structure introduced in Subsection 3.1, for a group of L frames, there exist $(L-1)!$ coding modes. The ideal way to find out the optimal $\mathbf{v}_{opt}$ in (2) is to loop over all the candidate coding modes and choose the combination that yields the minimal R-D cost in (1). While the complexity of this is intractably high, on the other hand, the interaction between the coding modes of different frames in a group prohibits optimizing each of them independently. To reduce the computational complexity, we use an *iterative descent algorithm* to find out the R-D optimized prediction modes for the group.

In this iterative approach, each time we minimize the Lagrangian cost $J_{\mathbf{v}}$ in (1) by varying one component in $\mathbf{v} = (v_0, v_1, ...v_{L-1})$, while keeping all other components constant. We repeat this for every frame in the group and obtain the minimal cost in this iteration. We repeat the iteration until the cost does not further decrease. The complete algorithm is described in Fig. 2. In our simulations, this process for a group of 10 frames usually takes no more than four iterations to converge. Hence the complexity reduces from $\mathcal{O}((L-1)!)$ to $\mathcal{O}(L)$.

4. SIMULATION RESULTS

We compare the performance of the proposed scheme (*prescient*) with a scheme that uses normal P-frames with periodic I-frame insertions to combat packet loss (referred to as the *P-I* scheme). Note that the P-I scheme intrinsically also provides certain amount of error-resilience. Also compared is the *greedy* scheme in [8]. We have implemented the three schemes by modifying the H.26L TML 8.5 [9]. The test sequences are *Foreman* and *Mother-Daughter*, representing high and moderate motion, respectively. 270 frames are coded, with the group size being $L = 10$ frames, and the frame rate is 30 fps. Coded frames are lost according to simulated conditions of channels with packet losses and varying delivery delay. The forward and backward (feedback) channel

0 Initialize $\mathbf{v} = (0, 1, 1, ..., 1)$; compute $R_{\mathbf{v}}$ according to (3); compute $\overline{D}_{\mathbf{v}}$ according to (4); $J_{\mathbf{v}}^{(0)} = \overline{D}_{\mathbf{v}} + \lambda R_{\mathbf{v}}$; $n = 1$.
1 Loop $i = 1 : L - 1$
2 Loop $v_i = 0 : i$
3 Compute $R_{\mathbf{v}}$ according to (3); compute $\overline{D}_{\mathbf{v}}$ according to (4); $J_{\mathbf{v}} = \overline{D}_{\mathbf{v}} + \lambda R_{\mathbf{v}}$.
4 End v_i
5 $v_{i,opt} = \arg\min_{v_i} J_{\mathbf{v}}$.
6 End i
7 $J_{\mathbf{v}}^{(n)} = \min J_{\mathbf{v}}$.
8 If $J_{\mathbf{v}}^{(n)} = J_{\mathbf{v}}^{(n-1)}$ stop; else $n = n + 1$ and go to 1.
9 $\mathbf{v}_{opt} = (v_{0,opt}, v_{1,opt}, ..., v_{L-1,opt})$.

Fig. 2. The iterative descent algorithm to determine the optimal coding modes.

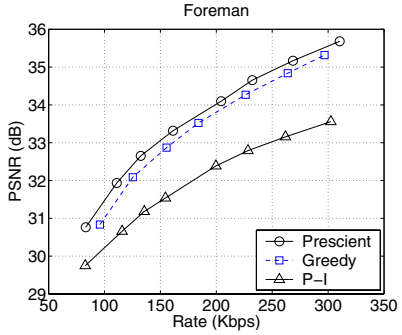

Fig. 3. R-D performance for *Foreman* sequence, $L = 10$.

have the same 2% packet loss rate and a delay distribution density modeled using a shifted Gamma distribution, which is specified by a shift of 25 ms, a mean of 100 ms, and a standard deviation of 50 ms. The playout deadline is 150 ms. The channel and streaming conditions above correspond to an effective overall loss rate of 10%, including the channel loss and the loss resulting from packets' late arrival. The PSNR of the decoded sequences is averaged over 15 random channel realizations. The first 20 frames of a sequence are not included in the statistics to exclude the influence of the transient period.

Fig. 3 shows the R-D performance for *Foreman* sequence. The distortion at different bitrates is obtained by varying the Q value and hence the Lagrange multiplier λ. Comparing the prescient scheme and the P-I scheme, a gain of 1.6 dB is observed at 200 Kbps and 2.0 dB at 300 Kbps, which corresponds to a bit rate saving of 40% at 33 dB. Note that although no retransmission is used, the video quality is still good over lossy channels. Fig. 4 shows the R-D performance for *Mother-Daughter* under the same experimental conditions. A gain of 1.3 dB is observed at 200 Kbps.

Comparing the prescient scheme with our previous greedy scheme, we observe typical gains of 0.4 dB for *Foreman* and 0.5 dB for *Mother-Daughter*. This corresponds to a bit rate saving of 11% at 31 dB for *Foreman*, and 12% at 34 dB for *Mother-Daughter*. The savings in bitrate, especially at lower rates, can be explained by the weaker error-protection applied for later frames in a group at lower rate cost. The complexity of the prescient scheme is also much lower for streaming pre-stored video.

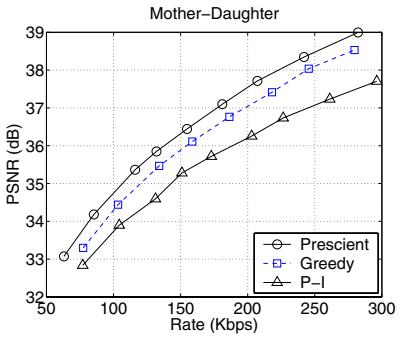

Fig. 4. R-D performance for *Mother-Daughter* sequence, $L = 10$.

5. CONCLUSIONS

We propose a prescient scheme to deliver pre-encoded video streams at very low-latency. We dynamically manage the prediction dependency for a group of packets by determining the prediction modes within a R-D optimization framework. The R-D optimized coding modes are determined by only considering the pre-computed R-D preamble instead of actually compressing the video, which greatly reduces the complexity at streaming time. Experiments demonstrate that the proposed scheme provides significant performance gains over simple Intra-insertion, and consistent bitrate savings compared to the greedy scheme.

6. REFERENCES

[1] P. A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," *IEEE Transactions on Multimedia*, Feb. 2001, submitted. http://research.microsoft.com/˜pachou.

[2] J. Chakareski, P.A. Chou, and B. Aazhang, "Computing rate-distortion optimized policies for streaming media to wireless clients," in *Proc. Data Compression Conference*, Apr. 2002, pp. 53–62.

[3] R. Zhang, S.L. Regunathan, and K. Rose, "End-to-end distortion estimation for RD-based robust delivery of pre-compressed video," in *Proc. 35th Asilomar Conf. Signals, Syst. and Computers*, Nov. 2001.

[4] T. Wiegand, N. Färber, and B. Girod, "Error-resilient video transmission using long-term memory motion-compensated prediction," *IEEE J. Select. Areas Commun.*, vol. 18, no. 6, pp. 1050–1062, June 2000.

[5] Yi J. Liang, M. Flierl, and B. Girod, "Low-latency video transmission over lossy packet networks using rate-distortion optimized reference picture selection," in *Proc. IEEE Int. Conf. on Image Processing (ICIP-2002)*, Rochester, NY, Sept. 2002, vol. 2, pp. 181 – 184.

[6] Rui Zhang, S.L. Regunathan, and K. Rose, "Prescient mode selection for robust video coding," in *Proc. IEEE Int. Conf. on Image Processing (ICIP-2001)*, Thessaloniki, Greece, Oct. 2001, vol. 1, pp. 974–7.

[7] Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, *Joint Final Committee Draft (JFCD) of Joint Video Specification (ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC)*, Aug. 2002, online available at: ftp://ftp.imtc-files.org/jvt-experts/.

[8] Yi J. Liang and B. Girod, "Low-latency streaming of pre-encoded video using channel-adaptive bitstream assembly," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, Lausanne, Switzerland, Aug. 2002, vol. 1, pp. 873 –876.

[9] ITU-T Video Coding Expert Group, *H.26L Test Model Long Term Number 8*, July 2001, online available at: ftp://standard.pictel.com/video-site/h26L/tml8.doc.

[10] Yi J. Liang, J.G. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: Does burst-length matter?" to appear, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, Apr. 2003.