LOW-LATENCY STREAMING OF PRE-ENCODED VIDEO USING CHANNEL-ADAPTIVE BITSTREAM ASSEMBLY

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

Today's Internet video streaming systems employ buffering and retransmission to guarantee the correct reception of each packet. This leads to high latency in media delivery. In this paper, we present an efficient low-latency Internet video streaming system that does not require retransmission of lost packets. We pre-store multiple representations of certain frames on the server such that a representation using a more reliable reference frame will be selected and sent during transmission. The potential mismatch error during bitstream assembly at transmission is avoided by using a novel layered coding structure with coding restrictions. The optimal picture type is determined within a rate-distortion framework during pre-encoding and transmission, adapting to the channel, so that the expected end-to-end distortion is minimized under the given rate constraint. The expected distortion is calculated based on an accurate binary tree modeling with the effects of channel loss and error concealment taken into account. Experiments demonstrate that the proposed scheme provides significant performance gains over a simple INTRA-insertion scheme. The increased error-resilience eliminates the need for retransmission, which allows to reduce the latency of streaming from 10-15 seconds to a few hundred milliseconds, with good video quality maintained.

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 desired but limited by the lossy nature of the transmission channel. For a typical hybrid video codec, an INTER frame is predicted from a reference frame with motion compensation, so that the temporal redundancy across successive frames is removed or reduced to provide higher coding efficiency. However, proper decoding of the INTER frame depends on the error-free reception and reconstruction of the reference frame 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 over the network and not recovered, the decoding of all subsequent frames depending on the lost frame will be affected. Hence, whenever a packet is lost, retransmission is required to guarantee the correct reception of each frame, which leads to higher latency in media delivery. Multiple retransmissions have to be made if the first retrial still fails, e.g., in the case of burst losses. The time for retransmissions constitutes the major part of the total end-to-end delay, and low latency streaming is therefore largely limited by packet loss and the motion compensation-based video coding.

In this work, in order to increase error-resilience and eliminate the need for retransmission, we pre-store multiple representations of certain frames at the server such that a representation can be chosen that only uses previous frames as reference that may be received with very high probability. We exploit the dependency across packets resulting from hybrid video coding and dynamically manage this dependency to achieve increased error-resilience. If the need for retransmission can be eliminated, buffering is needed only to absorb the packet delay jitter, so that the buffering time can be reduced to a few hundred milliseconds.

For on-demand streaming, pre-compressing the video offline greatly decreases the complexity of the server. During media delivery, the server only has to dynamically assemble the appropriate bitstream while adapting to the channel condition. This can be easily achieved in real-time with very low latency since no encoding or decoding is involved during service. Storing multiple versions of the video on the server requires relatively higher disk storage. However, the past few years have seen a significant drop of the prices of disk storage. The increased storage overhead is hence more and more affordable and obviously worthwhile when considering the benefit it brings to the large number of users who may watch the same video.

A major challenge for streaming pre-stored video is the potential mismatch error resulting from assembling the bitstream. Usually, when a particular representation of a frame is selected and sent, the reference available for decoding that frame at the decoder might be different from the one used for encoding. This mismatch error might propagate and lead to severe degradation in performance. Past work addressing this problem includes using S-frames [1], and SP-frames, as proposed for H.26L [2]. However, both are achieved high coding costs. In this work, we avoid the mismatch error by using a layered coding structure that allows switching between pre-coded streams at restricted positions.

This paper is structured as follows: we first describe dynamic management of packet dependency. In Section 3, we present the layered coding structure and optimal picture type selection. Experimental results are presented in Section 4.

This work has been supported by a gift from HP Labs, Palo Alto, CA, and by the Stanford Networking Research Center.

2. DYNAMIC MANAGEMENT OF PACKET DEPENDENCY FOR ERROR-RESILIENCE

In a conventional coding and transmission scheme without any awareness of channel losses, an I-frame is typically followed by a series of P-frames (referred to as *P1-frames* in this work), 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 error [3]. We refer to this type of P-frame obtained as *P2-frame*, and further extend the types of INTER coded pictures to *Pv-frame*, where $v = 1, 2, 3, ... V$ denotes which reference frame is used for prediction and indicates prediction dependency. Using different types of pictures other than the normal P1-frames increases error-resilience. The robustness is normally obtained at the expense of a higher bitrate since the correlation between two frames becomes weaker in general as they are more widely separated.

An earlier related proposal is the Reference Picture Selection mode (RPS) in Annex N of H.263+ to terminate error propagation based on feedback [4], [5]. When the encoder learns through the feedback channel that a previous frame is lost, instead of using the most recent frame as a reference, it can code the next P-frame based on an older frame that is known to be correctly received at the decoder [6], [7]. The multiframe prediction support in Annex N was later subsumed by the more advanced Annex U of H.263++ and is now an integral part of the emerging H.26L standard [2]. In our work [3], we extend the RPS concept by allowing the use of a reference frame whose reception status is uncertain but whose reliability can be inferred, for live-encoding. This work extends [3] and focuses on the streaming of pre-encoded video.

In [8], long-term memory (LTM) prediction is used for both improved coding efficiency and error-resilience over wireless channels. Different macroblocks in a frame may be predicted from different reference frames, which makes it difficult to put an entire frame into an IP packet and manage the prediction dependency at the packet level during transmission. In this work, we avoid this problem by selecting the reference at the frame 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 Picture Type Selection (OPTS)* within a rate-distortion (RD) framework, by considering video content, channel loss probability and channel feedback (e.g., ACK, NACK, or time-out). This applies to both pre-encoding the video offline and assembling the bitstreams during streaming. When coding a frame, several trials are made, including using the I-frame as well as INTER coded frames using different reference frames in the long-term memory, e.g., $v = 1, 2, 3, ... V$. V is the number of previous decoded frames available from the LTM, referred to as the *length of LTM*. The associated rate and expected distortion are obtained to calculate the cost for a particular trial through a Lagrangian formulation. The expected distortion is the statistical average of the distortion of all possible decoded outcomes. The distortions are obtained through an accurate binary tree modeling, the details of which will be discussed in Section 3.2. The optimal picture type v_{opt} is selected such that the minimal RD cost is achieved.

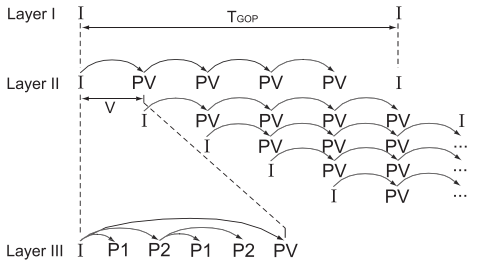

Fig. 1. Layered coding structure with coding restrictions. $T_{GOP} = 25, V = 5$.

3. LAYERED CODING STRUCTURE USING RESTRICTED OPTS FOR BITSTREAM ASSEMBLY

We avoid any mismatch error that might occur during bitstream assembly at transmission by using a layered coding structure with coding restrictions on each layer and by providing multiple representations of certain key frames on the server.

3.1. Layered coding structure with coding restrictions

The coding structure consists of three layers as shown in Fig. 1:

1) *Layer I* pictures are I-frames that serve as the lead of a *group of pictures* (GOP). The maximum length of a GOP is T_{GOP}, a multiple of the LTM length V.

2) *Layer II* pictures serve as the lead of a *sub-group of pictures* (SGOP), and only two types of pictures are allowed in this layer: I-frames and PV-frames. Layer I and Layer II pictures are also referred to as *SYNC-frames*, which are only positioned at kV, where $k = 0, 1, 2, ...$ (a video sequence starts from 0), with a fixed interval of V. The prediction dependency of SYNC-frames in a GOP is shown in Fig. 1. T_{GOP}/V versions of SYNC-frames in different phases are pre-stored, with the leading I-frame starting from different positions as shown in Fig. 1. The deployment of the SYNC-frames with multiple phases allows the use of an I-frame during assembly at any SYNC positions (kV) without resulting in any mismatch error. The choice between keeping using a PV-frame in a GOP or switching to an I-frame is determined within the OPTS framework considering channel conditions and feedback, which will be discussed in the next subsection. Once an I-frame is inserted, it starts a new GOP, and the next I-frame has to be inserted in T_{GOP} frames, if not sooner.

The SYNC-frames of Layer I and II also facilitate interactive functions such as random access, fast-forward or fast-reverse.

3) *Layer III* pictures are those within a SGOP following its lead. The Layer III picture can only be either an I-frame (note that this is not a SYNC-frame though) or a Pv-frame that uses a previous frame in the same SGOP (including the SGOP-lead) as a reference. This limits the prediction dependency within a SGOP, and the picture type is also selected in the OPTS framework. Since the SGOP-lead has T_{GOP}/V possible versions, T_{GOP}/V sets of corresponding Layer III pictures are needed to avoid any mismatch error. In summary, T_{GOP}/V representations of each frame have to be pre-stored in this scheme. Alternatively, if storage is more of a concern, fewer versions of Layer III pictures may be pre-stored. This leads to a mismatch error when the SGOP-lead sent is not the one used for pre-encoding. However the mismatch is mitigated

since it is limited within a particular SGOP and will not propagate to the next.

Layer III frames are pre-coded offline based on the video content and the expected channel loss rate, while SYNC-frames are dynamically assembled during delivery to respond to the channel feedback. The SGOP-lead determines which set of Layer III pictures to send, if more than one are pre-coded. Pre-stored multiple versions of the bitstreams enable mismatch-free assembly, though achieved at a moderate cost of extra disk storage. The layered coding also provides temporal scalability for rate control. Pictures in different layers are coded with certain restrictions. However, we will show in Section 4 that, over lossy channels, the added restrictions do not result in much performance degradation compared to live-encoding, in which frames are coded during transmission and without any restriction in selecting the reference frame.

3.2. Optimal picture type selection using binary tree modeling

We begin with explaining the coding of Layer III pictures. A binary tree, shown in Fig. 2, is used to represent error propagation from frame to frame and all possible decoded outcomes at the decoder. A *node* in the tree represents a possible decoded outcome (frame) at the decoder. In the example in Fig. 2, Frame kV, e.g., a SGOP-lead, has only one node with probability 1. Frames $kV+1$ and $kV+2$ are, for instance, both P1-frames. Two *branches* leave the node of Frame kV representing two cases that either reference frame kV is properly received (and decoded) with probability q or lost with probability $p = 1 - q$. These two cases lead to two different decoded versions of Frame $kV+1$, provided that Frame $kV+1$ is available at the decoder. The upper node of Frame $kV+1$ is obtained by normal decoding process using the correct reference (decoded kV); and the lower node corresponds to the case if Frame kV is lost. In the latter case, a simple concealment is done by copying Frame $kV-1$ to kV, and Frame $kV+1$ hence has to be decoded using the concealed reference instead. This leads to a mismatch error that might propagate at the decoder, depending on the prediction dependency of the following frames. The distortion associated with these two cases are evaluated by decoding $kV+1$ at the *encoder* side. The value of q or p is estimated from the accumulated channel statistics, which may be updated as the channel conditions vary.

While encoding the next frame $kV+3$, several trials are made using different picture types. We use $v(kV+n)$ to denote the reference frame that Frame $kV+n$, the n-th frame in a SGOP, may use. For a particular $v(kV+n)=v$, a rate R_v is obtained and the expected distortion of all decoded outcomes is given by

$$\overline{D}_v = \sum_{l=1}^{L(kV+n)} p_{vl} D_{vl}, \qquad (1)$$

where $L(kV+n)$ is the number of nodes for Frame $kV+n$. p_{vl} is the probability of outcome (node) l, which can be calculated easily from the tree model if statistical independence of successive losses is assumed. D_{vl} is the distortion associated with the decoded outcome l. Note that D_{vl} includes both the quantization error and possible decoding mismatch error, which is calculated accurately at the encoder side.

With the obtained rate and expected distortion, the Lagrangian cost associated with using a particular Pv-frame is

$$J_v = \overline{D}_v + \lambda R_v. \qquad (2)$$

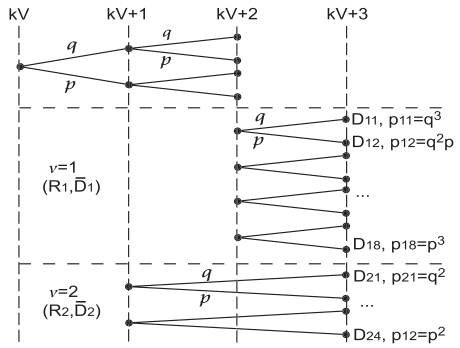

Fig. 2. The binary tree for the estimate of error propagation and optimal picture type selection.

Comparing all candidate reference frames, $v(kV+n) = 1, 2, \ldots n$ and INTRA coding (denoted by $v=\infty$), the optimal picture type $v_{opt}(kV+n)$ is the one that results in minimal J_v

$$v_{opt}(kV+n) = \arg\min_{v=1,\,2,\,\ldots\,n,\,\infty} J_v(kV+n). \qquad (3)$$

In (2), 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 [2], and Q is the quantization parameter used to trade off rate and distortion.

Different from the approach in [8], we keep tracking the states of each frame by storing all possible decoded outcomes in the LTM of the encoder. For each frame to be encoded, all of its possible decoding outcomes can be obtained using the saved outcomes of its reference frame (either correct or concealed) from the LTM. Comparing to a recursive estimation of the decoder distortion in [9], which is only precise for integer pixels, the proposed binary tree modeling is also accurate for half-pixel, quarter-pixel or one-eighth-pixel precision. An analysis of the storage complexity of this model can be found in [3].

The binary tree model is also applied to the coding of the SYNC-frames of Layer I and II, where only two types of pictures are allowed to use. The obtained costs are saved for different versions such that the optimal picture type is dynamically selected at assembly, without any need for decoding. We require that the LTM length V is greater than or equal to the feedback delay in frames, e.g., when sending a SYNC-frame, the reception status of the previous SYNC-frame is known. This ensures that the impairment of a possible SYNC-frame loss can be limited within one SGOP.

4. SIMULATION RESULTS

We compare the performance of three channel-adaptive schemes: 1) the proposed OPTS scheme for pre-encoding and transmission, where V versions of each coded frame is stored; 2) a scheme that uses normal P1-frames with periodic I-frames to combat packet loss (referred to as the *P-I* scheme). V phases of the sequence are pre-encoded, which is similar to the multi-phase coding structure of Layer II in the OPTS scheme. This allows feedback-induced I-frame insertion at certain positions without any mismatch error. Note that the P-I scheme intrinsically also provides certain amount of error-resilience. 3) A channel-adaptive live-encoding scheme using OPTS, where the sequence is encoded at transmission, and each frame may choose any reference frame available from the

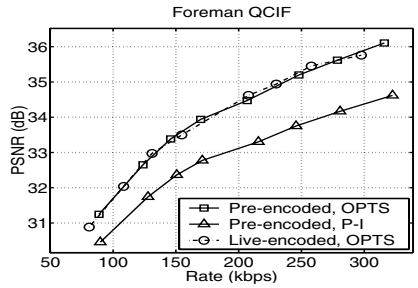

Fig. 3. RD performance for *Foreman* sequence, 10% packet loss, 5 frames feedback delay.

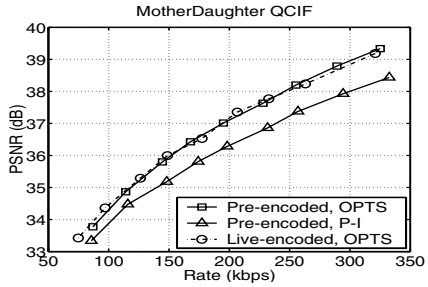

Fig. 4. RD performance for *Mother-Daughter* sequence, 10% packet loss, 5 frames feedback delay.

LTM with no restriction. This scheme serves as a baseline for comparison and cannot be implemented for pre-encoding. We have implemented the three schemes by modifying the H.26L TML 8.5 [2]. The test sequences are *Foreman* and *Mother-Daughter*, representing high and moderate motion, respectively. 230 frames are coded, and the frame rate is 30 fps. Coded frames are dropped according to simulated channel conditions with a range of loss probabilities. The PSNR of the decoded sequences is averaged over 30 random channel loss patterns. The first 30 frames of a sequence are not included in the statistics to exclude the influence of the transient period.

Fig. 3 shows the RD performance of sending the *Foreman* sequence over the channel with 10% loss rate. Feedback delay is 5 frames, and the length of LTM V is 5. The distortion at different rates is obtained by varying the Q value and hence the Lagrange multiplier λ. Comparing schemes 1 and 2, a gain of 1.2 dB is observed at 200 kbps and 1.5 dB at 300 kbps by using the proposed scheme, which corresponds to a bit rate saving of 36% at 34 dB. The gain is typically higher at higher rates since at lower rates, LTM prediction with $v > 1$ is less efficient and the advantage of OPTS decreases. Note that although no retransmission is used, the video quality is still good over lossy channels. Fig. 4 shows the RD performance of *Mother-Daughter* under the same experimental conditions. A gain of 0.5 dB is observed at 200 kbps and 0.9 dB at 300 kbps. The gain from using the proposed scheme is lower compared to *Foreman* since the impairment of frame loss is smaller due to lower motion in the sequence.

Distortion at different channel loss rates is shown in Fig. 5 for *Foreman* encoded at approximately the same 200 kbps using the three schemes. The gain is observed ranging from 0.5 dB to 1.6 dB, depending on the channel loss rate. The advantage of using error-resilient OPTS is more obvious at higher loss rate. From

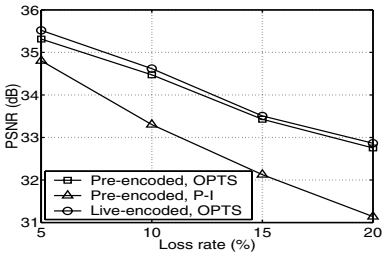

Fig. 5. Distortion at different channel loss rates. *Foreman* sequence. The bitrate is 200 kbps.

Fig. 3-5 it is also observed that the performance gap between pre-encoding and live-encoding is no more than 0.2 dB at various channel loss rates, which indicates that the restrictions in the pre-encoding do not result in much performance degradation.

5. CONCLUSIONS

We propose a channel-adaptive video streaming scheme that dynamically manages the dependency across packets and selects the picture type that is optimal within an RD framework. Experiments with packet loss rates between 5% and 20% demonstrate significant gain. The potential mismatch error during bitstream assembly is avoided by using a layered coding structure and by providing multiple versions of the coded sequence on the server. The increased error-resilience eliminates the need for retransmission, which allows to reduce the latency for Internet video streaming to a few hundred milliseconds, with good video quality maintained.

6. REFERENCES

[1] N. Färber and B. Girod, "Robust H.263 compatible video transmission for mobile access to video servers," in *Proc. IEEE Int. Conf. on Image Processing*, Oct. 1997, vol. 2, pp. 73–76.

[2] 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.

[3] Y. 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)*, Sept. 2002, Rochester, NY.

[4] ITU-T Recommendation H.263 Version 2 (H.263+), *Video coding for low bitrate communication*, Jan. 1998.

[5] S. Wenger, G.D. Knorr, J. Ott, and F. Kossentini, "Error resilience support in H.263+," *IEEE Journal on Circuits and Systems for Video Technology*, vol. 8, no. 7, pp. 867–877, Nov. 1998.

[6] B. Girod and N. Färber, "Feedback-based error control for mobile video transmission," *Proceedings of the IEEE*, vol. 87, no. 10, pp. 1707 – 1723, Oct. 1999.

[7] S. Lin, S. Mao, Y. Wang, and S. Panwar, "A reference picture selection scheme for video transmission over ad-hoc networks using multiple paths," in *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, Aug. 2001.

[8] 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.

[9] R. Zhang, S.L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE J. Select. Areas Commun.*, vol. 18, no. 6, pp. 966 –976, June 2000.