LOW-LATENCY VIDEO TRANSMISSION OVER LOSSY PACKET NETWORKS USING RATE-DISTORTION OPTIMIZED REFERENCE PICTURE SELECTION

Yi J. Liang, Markus Flierl and Bernd Girod

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

ABSTRACT

Today's Internet video streaming systems employ buffering and multiple retransmissions to guarantee the correct reception of each packet, which leads to high latency in media delivery. In this work, a novel scheme of dynamic management of the dependency across packets is proposed using optimal reference picture selection at the frame level that adapts to the channel. The optimal selection of the reference picture is achieved with a rate-distortion framework, which minimizes the expected end-to-end distortion given a 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 of retransmission, which makes it possible to reduce the latency from 10-15 seconds to a few hundred milliseconds, with similar video quality maintained.

1. INTRODUCTION

Today's Internet provides best-effort services without any guarantee of quality. Internet video streaming has to cope with this lack of QoS guarantees. Due to congestion and the heterogeneous infrastructure, the transmission is plagued by variability in throughput, packet loss, and delay. To mitigate these effects, today's 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 interactivity of playback, such as the VCR functionality.

Low latency in video transmission is desired but limited by the lossy nature of the transmission channel. In a typical hybrid video codec, an INTER frame is predicted from a reference frame with motion compensation, so that the temporal redundancy across adjacent frames is removed or reduced to provide higher coding efficiency. However, proper decoding of the INTER frame depends on the correct 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 channel and not well 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 every 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 errors. 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 way video sequences are coded.

In this work, in order to increase error-resilience and eliminate the need for retransmission, we consider the dependency across packets as a result of hybrid video coding, and dynamically manage this dependency while adapting to the channel condition. If the need for retransmission can be eliminated, buffering is needed only to absorb the packet delay jitter, so that buffering time can be reduced to a few hundred milliseconds.

An earlier proposal is the Reference Picture Selection mode (RPS) proposed in Annex N of H.263+ to terminate error propagation based on feedback [1], [2]. 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 [3], [4]. 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 [5]. In this work, we extend the RPS concept by allowing the use of a reference frame whose reception status is uncertain but whose reliability can be inferred. This selection is optimal within an end-to-end RD framework and channel-adaptive.

In [6], long-term memory (LTM) prediction is used for both improved coding efficiency and error-resilience over wireless channels. Different macroblocks in a same frame may be predicted from different reference frames, which makes it difficult to put a frame into an IP packet and manage the dependency at the packet level. In this work we avoid this problem by selecting the reference at the frame level. In [7], the decoder distortion is recursively estimated for optimal selection of the INTER/INTRA mode for each macroblock, which is only precise for integer pixels. In this work, the estimation using the proposed binary tree modeling is also accurate for half-pixel, quarter-pixel or one-eighth-pixel precision.

2. CHANNEL-ADAPTIVE PACKET DEPENDENCY MANAGEMENT

In a conventional encoding and transmission scheme without any awareness of channel losses, an I-frame is typically followed by a series of P-frames, which are predicted from their immediate predecessors. This scheme is vulnerable to channel errors since each P-frame depends on its predecessor and any packet loss will break

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

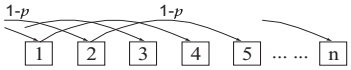

Fig. 1. A coding structure where each frame uses the third previous frame as a reference ($v = 3$). Assuming enough frames before the Frame 1 are available for reference. Each frame is correctly received at the decoder with probability $1 - p$. Frame 5 in the sequence depends on $\lceil \frac{5}{3} \rceil = 2$ previous frames, and the probability it will be affected by a previous loss is $p_e = 1 - (1 - p)^2$.

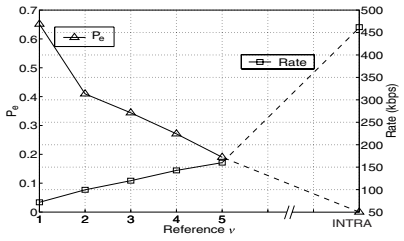

Fig. 2. The probability of the 10th frame being affected by a prior loss (left axis) and the sequence-average rates (right axis) using different reference frames. Rates are obained by encoding the first 230 frames of *Foreman* sequence (30 frame/sec) using H.26L TML 8.5 at an average PSNR of approximately 33.4 dB. $p = 0.10$.

the prediction chain and affect all subsequent P-frames. If each P-frame is predicted from the frame preceding the previous frame instead, the scheme is more robust against channel errors due to the changed dependency. Consider, for example, a fixed coding structure where each frame uses the reference that is v frames back for prediction. The n-th frame in the sequence thus depends on $\lceil \frac{n}{v} \rceil$ previous frames, where $\lceil x \rceil$ represents the smallest integer number that is greater than or equal to x. An example of $v = 3$ is illustrated in Fig. 1 . Assuming each packet is lost independently with probability p, the probability that the n-th frame in the sequence will be affected by a previous loss is hence

$$p_e = 1 - (1 - p)^{\lceil \frac{n}{v} \rceil}. \quad (1)$$

This probability is plotted in Fig. 2 for $p = 0.10$, $n = 10$, and $v = 1, 2, ... 5$, and INTRA coding (we use $v = \infty$ to denote INTRA coding).

As illustrated in Fig. 2, using frames from the long-term memory with $v > 1$ for prediction, instead of using an immediately previous frame ($v = 1$), reduces prediction efficiency and 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. A special and extreme case is the I-frame, which is the most robust over lossy channels, but generally requires 5-10 times as many bits as the P-frame. In Fig. 2, we also show the average rates of encoding the *Foreman* sequence at close PSNRs using different reference v and INTRA coding.

Fixed reference selection schemes provide different amount of error resilience at different coding costs, as is shown in Fig. 2. In the example illustrated in Fig. 3, we show the significant gains by optimal reference picture selection (referred to as the *ORPS* scheme) and dynamic packet dependency management according to channel conditions. The reference used by each frame and the

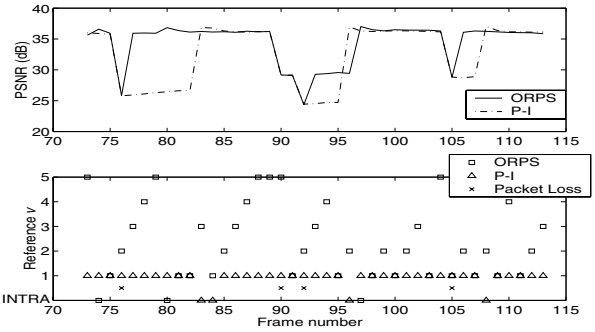

Fig. 3. Frame PSNR of part of the decoded *Foreman* sequence and the reference frame used by each frame. Channel loss rate is 10%. LTM length is 5 frames, and channel feedback delay is 7 frames. The complete sequence includes the first 230 frames of *Foreman* and the rates for the ORPS and the P-I scheme are 201.1 kbps and 201.8 kbps respectively; the average decoded PSNRs are 34.26 dB and 33.21 dB respectively.

decoded PSNR are plotted in Fig. 3. Our model bases its selection on channel loss rate and feedback information. The scheme for comparison uses normal P-frames ($v = 1$) with periodic I-frame insertion plus NACK-triggered I-frame insertion (referred to as the *P-I* scheme) to combat packet loss. At approximately the same rates, the proposed scheme gains more than 1 dB on average. Since packet dependency is optimally manipulated, INTER coded frames are more resilient to packet loss. E.g., when the 76th packet is lost, the error, which could otherwise propagate through Frame 82, does not affect any other frame since it is not used as a reference. Note that with the proposed scheme, the video quality is good without any retransmission of lost packets.

3. OPTIMAL REFERENCE PICTURE SELECTION

Due to the trade-off between error-resilience and coding efficiency, we take the channel loss into consideration and select the reference picture within a rate-distortion (RD) framework. We use the binary tree structure to represent error propagation from frame to frame and all possible decoded outcomes at the decoder.

Fig. 4 illustrates such a tree structure. A *node* in the tree represents a possible decoded outcome (frame) at the decoder. In the example shown in Fig. 4, Frame $n - 3$ has only one node with probability 1 (e.g., due to the reception status confirmed by feedback). Frames $n-2$ and $n-1$ both, for instance, use their immediately preceding frames as references. Two *branches* leave the node of Frame $n - 3$ representing two cases that either reference frame $n - 3$ is properly received (and decoded) with probability q or lost with probability $p = 1 - q$. These two cases lead to two different versions of Frame $n - 2$, provided that Frame $n - 2$ is available at the decoder. The upper node of Frame $n-2$ is obtained by normal decoding process using the correct reference (decoded $n - 3$); and the lower node corresponds to the case when Frame $n - 3$ is lost. In the latter case, a simple concealment is done by copying $n - 4$ to $n - 3$, and Frame $n - 2$ hence has to be decoded using the concealed reference (decoded $n - 4$). This leads to the 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 $n - 2$

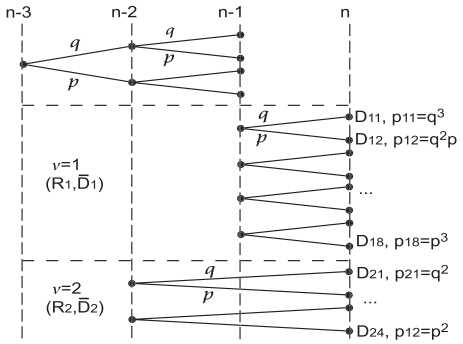

Fig. 4. The binary tree structure for the estimate of error propagation and optimal reference selection.

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 change.

In encoding the next frame n, several trials are made using different reference frames in the LTM and the resulting rates and expected distortion are obtained in order to calculate the cost function. Assuming V previously decoded frames are available from the long-term memory (V is referred to as the *length of LTM*), we use $v(n)$ to represent the reference frame that Frame n may use. For example $v(n) = 1$ denotes using the previous frame and $v(n) = 2$ denotes using the frame preceding that frame, and so on. For a particular $v(n) = v$, a rate R_v is obtained and the expected distortion of all decoded outcomes is given by

$$\overline{D}_v = \sum_{l=1}^{L(n)} p_{vl} D_{vl}, \tag{2}$$

where $L(n)$ is the number of nodes for Frame n, and $L(n) = 2L(n - v(n))$. p_{vl} is the probability of outcome (node) l, which can be calculated easily from the tree structure, if statistical independence of successive losses is assumed. For example, in Fig. 4, $p_{11} = q^3$, while $p_{12} = q^2 p$ and so on. 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. For simplicity we assume one frame is contained in one IP packet in this work. If a frame is large enough and split into more than one packet, the frame can still be represented by a node in this model, but the outcome probabilities will slightly change in the tree structure.

With the obtained rate and expected distortion, the Lagrangian cost associated with using the reference frame $v(n) = v$ is

$$J_v = \overline{D}_v + \lambda R_v. \tag{3}$$

Comparing all candidate reference frames, $v(n) = 1, 2, ... V$ and ∞ (INTRA coding), the optimal reference frame $v_{opt}(n)$ is the one that results in minimal J_v

$$v_{opt}(n) = \arg\min_{v=1,\,2,\,\ldots\,V,\,\infty} J_v(n). \tag{4}$$

In (3), λ is a Lagrange multiplier and 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 [5], [8], and Q is the quantization parameter used to trade off rate and distortion. Note that in each stage $v_{opt}(n)$ is obtained given the condition that frame n is available at the decoder, which means the reception status of n is not considered in selecting the reference frame for n at the encoder.

In [6], a similar binary tree structure is used to select the optimal reference for each macroblock. However only three branches are considered in the binary tree to calculate the approximate expected distortion, and a second Lagrange multiplier κ is used to trade off error-resilience and coding efficiency. An approximate model has to be used in [6] because of several difficulties mentioned: (1) tree size grows at the rate of $L = 2^n$ for each macroblock and modeling can be soon intractable; (2) time instants do not correspond to levels of the tree due to LTM prediction.

In this work, we take the advantage of channel feedback in order to limit the tree size, which is reasonable and necessary in practice. When the reception status is known (e.g. from ACK, NACK, or time-out) for a previously sent packet, half of the branches leaving the corresponding node are removed, and so are all the dependant nodes. In this way, the maximum size of the tree is kept constant as the states propagate. In general, given the feedback delay d_{fb} in frames (when encoding Frame n, the status of Frame $n - d_{fb}$ becomes known), the maximum number of outcomes kept for a frame is $L = 2^{d_{fb}-1}$. Note that the reference selection algorithm itself does not necessarily depend on any feedback.

In solving the second difficulty mentioned above, 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 are obtained using the saved outcomes of the reference frame (either correct or concealed) from the LTM. Therefore the binary tree modeling here is accurate in estimating distortions. This is achieved at higher storage complexity, depending on the length of LTM, V, and the feedback delay d_{fb}. Noting that the maximum decoded frames kept for a most recent frame encoded is $L = 2^{d_{fb}-1}$, the maximum number of decoded frames that have to be stored in the LTM is

$$\begin{cases} \sum_{k=d_{fb}-1-V}^{d_{fb}-1} 2^k & \text{for} \quad V \le d_{fb} - 1; \\ [V - (d_{fb} - 1)] + \sum_{k=0}^{d_{fb}-1} 2^k & \text{for} \quad V > d_{fb} - 1. \end{cases} \tag{5}$$

For instance, when using $V = 5$ and $d_{fb} = 7$, the maximum number of decoded frames that have to be stored is 126, which corresponds to about $4.6MB$ $(1.5 \times 176 \times 144 \times 126/1024^2)$ for QCIF sequences. The increased memory overhead is obviously affordable and worthwhile for the media server when considering the gain in error-resilience and low latency.

4. SIMULATION RESULTS AND PERFORMANCE COMPARISON

We compare the performance of the proposed optimal reference picture selection (*ORPS*) scheme with that of the *P-I* scheme. Note that the P-I scheme intrinsically also provides certain amount of error resilience. We have implemented the two schemes by modifying the H.26L TML 8.5. The video sequences used are *Foreman* and *Mother-Daughter*, representing two extremes in terms of the amount of motion. 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 and no retransmission is used. The PSNR of the decoded sequences is averaged over 30 random channel loss patterns. The first 30 frames are not included in the statistics to exclude the influence of the transient period.

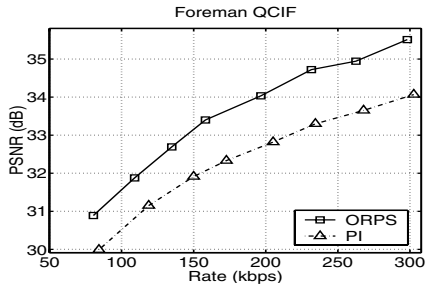

Fig. 5. RD performance of *Foreman* sequence. $V = 5$, $d_{fb} = 7$, $p = 0.10$.

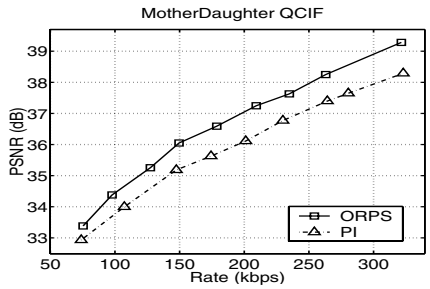

Fig. 6. RD performance of *Mother-Daughter* sequence. $V = 5$, $d_{fb} = 7$, $p = 0.10$.

Fig. 5 shows the RD performance of transmitting the *Foreman* sequence over the channel with 10% loss rate. Feedback delay is 7 frames, and the length of LTM is 5. The distortion at different rates is obtained by varying the Q value and hence the Lagrange multiplier λ. The INTRA rate in the P-I scheme is adjusted such that the rates keep up with the ORPS scheme at close Q values. A gain of 1.2 dB is observed at 200 kbps and 1.5 dB at 300 kbps by using the ORPS scheme, which corresponds to a bit rate saving of 35% 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 ORPS decreases. Fig. 6 shows the RD performance of *Mother-Daughter* sequence under the same experimental conditions. A gain of 0.9 dB is observed at 200 kbps and 1.0 dB at 300 kbps. The gain from using ORPS is lower compared to *Foreman* sequence since the effect of frame loss is smaller due to lower motion in the sequence.

Distortion at different channel loss rates is shown in Fig. 7 for *Foreman* encoded at approximately the same 200 kbps using the two schemes. The gain is observed ranging from 0.7 dB to 1.8 dB, depending on the channel loss rate. The advantage of using error-resilient ORPS is more obvious at higher channel loss rate. The RD performance with different LTM length V is shown in Fig. 8. At a feedback delay of 7 frames, an increase of V from 2 to 5 results in 0.5-0.7 dB gain at higher rates while an increase from 7 ($= d_{fb}$) to 8 does not give much further improvement. This gives us some idea on how to choose the LTM length for the trade-off between performance and storage complexity.

5. CONCLUSIONS

We propose a channel-adaptive video coding scheme that dynamically manages the dependency across packets and selects the reference at the frame level that is optimal within an RD framework. Experiments with packet loss rates between 5% and 20% demonstrate significant gain. The increased error resilience eliminates the need for retransmission, which allows to reduce the latency for Internet video streaming to a few hundred millisecond with the video quality well maintained.

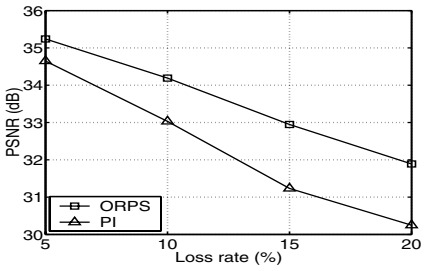

Fig. 7. Distortion at different channel loss rates. *Foreman* sequence. $V = 5$, $d_{fb} = 7$.

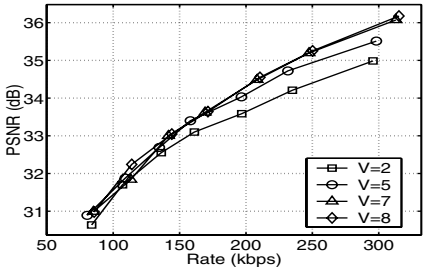

Fig. 8. RD performance at different LTM length. *Foreman* sequence. $d_{fb} = 7$, $p = 0.10$.

6. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control," in *Proc. IEEE Int. Conf. on Image Processing*, Oct. 2001, vol. 3, pp. 542–545, Thessaloniki, Greece.