RECEIVER ERROR CONCEALMENT USING ACKNOWLEDGE PREVIEW (RECAP) – AN APPROACH TO RESILIENT VIDEO STREAMING

Chuohao Yeo

University of California, Berkeley
Dept. of EECS
Berkeley, CA 94720, USA

Wai-Tian Tan, Debargha Mukherjee

Hewlett-Packard Laboratories, Palo Alto
MCNL
Palo Alto, CA 94304, USA

ABSTRACT

High-quality and low-latency video streaming is essential to providing a natural user experience in video conferencing. This is challenging over lossy networks since compressed video is highly fragile while the low-latency requirement limits the effectiveness of traditional error control approaches such as retransmission and forward error correction. In this paper, we advocate a practical solution for low-latency video communications over best-effort networks that employs an additional low-quality, low-resolution but robustly coded copy of the video. This approach, called RECAP, incurs minimal rate overhead, and can be combined with previously decoded frames to achieve effective concealment of isolated and burst losses even under tight delay constraints. RECAP achieves PSNR gains of 2-6 dB against complete frame loss.

Index Terms— video streaming, low-latency, error resilient, error concealment, high-definition video

1. INTRODUCTION

Supporting video conferencing experience that is natural and collaborative is challenging over best-effort networks. First, higher definition sources, such as 720p, that stress networks are often necessary for effective face-to-face communication. Second, occasional picture break-ups and freezes that plague Internet streaming are unacceptable. Fundamental to satisfying these requirements is an effective error control scheme that operates with low-latency and low bit-rate overhead.

Error resilient video streaming over unreliable networks has been well-studied [1]. One general approach that is widely practiced is retransmission of any lost data. While effective, retransmissions nevertheless increase latency. Given a round-trip time (RTT) of over 200 ms between California and Singapore, a mere 1 or 2 retransmissions would cause communications to appear non-interactive.

Rather than reacting to losses, Forward Error Correction (FEC) is another general approach that avoids retransmissions by proactively transmitting parity data. A video specific variant of FEC is Unequal Error Protection (UEP) that preferentially provides more protection to important data such as macroblock modes and motion vectors. For both FEC and UEP, interleaving (that increases latency) is typically necessary for channels with bursty losses. When the time horizon for protection is 1 video frame, FEC/UEP is powerless against loss bursts that last one or more frames.

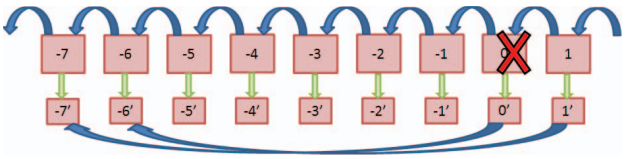


Fig. 1. **RECAP basics:** An independent preview layer that only predicts from positively acknowledged frames ensures proper decoding whenever a preview frame is received. This acknowledged preview serves as "ground truth" for effective receiver error concealment under both isolated loss and burst loss. In above, frame 0 is lost, and its preview (0') helps concealment of frame 0. Even if frames -2, -2', -1, -1', 0 and 0' are all lost, a received 1' is still correctly decoded and can be used for immediate concealment of 1.

Instead of attempting lossless recovery of data, receiver error concealment is a third general approach that accepts data loss and attempts to intelligently estimate the missing data. Generally, it is more difficult to conceal a single large and contiguous region than multiple small and scattered regions. Towards this end, H.264 supports Flexible Macroblock Ordering (FMO) which allows macroblocks in slices to be arranged in a checker-board pattern for more effective concealment should one slice be lost. Nevertheless, FMO does not provide any added protection if all slices in a frame are lost.

Reference Picture Selection (RPS) is a feature in H.263v2 and H.264 that does not increase latency and has proved to be effective against losses for video conferencing. Error propagation is stopped when an encoder references only past frames that has been positively acknowledged. Nevertheless, RPS is a reactive scheme with reaction time of a single RTT. For a RTT of over 200 ms, a loss could translate into 6 or more corrupted video frames.

In this work, we present an error control scheme named *Receiver Error Concealment using Acknowledge Preview* (RECAP) that improves upon RPS such that visual quality can be high even when round-trip delay is large. As shown

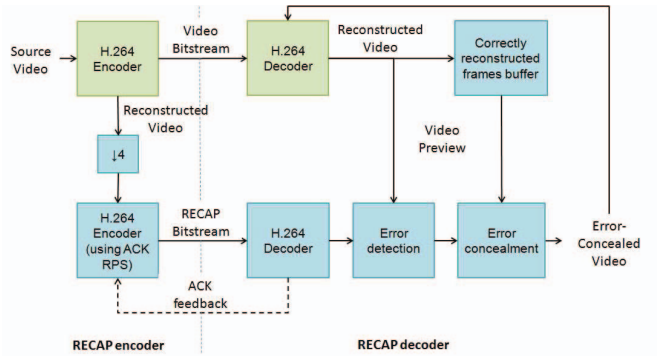


Fig. 2. **RECAP encoder and decoder block diagrams.**

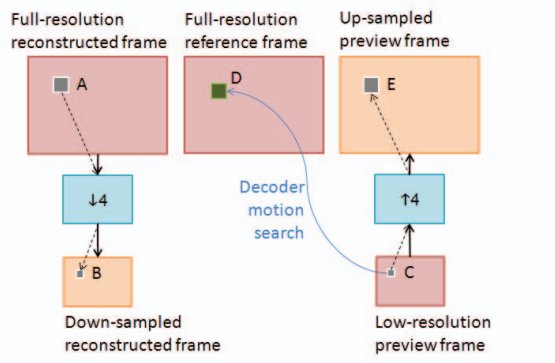


Fig. 3. **Illustration of RECAP error detection and concealment.** To test if a concealed block A is consistent with the video preview, RECAP computes and thresholds the MSE between its down-sampled version, block B, and the co-located block from the low-resolution frame, block C. If it is found to be inconsistent, block C is then used to perform decoder motion search in the smoothed full-resolution reference frame. For example, block D could be the best match found. A choice is then made between block D and the up-sampled block E for the final concealment.

in Fig. 1, RECAP employs an additional and independent low-resolution video stream, or *preview*. For compression efficiency, the main presentation normally predicts from any past pictures and is restricted to predict from a positively acknowledged picture only when loss is detected. In contrast, the preview stream *always* predicts from a positively acknowledged frame. The key advantage of this *acknowledged preview* is that every received preview picture can be properly decoded. Furthermore, when the main presentation is lost, *e.g.* frame 0 in Fig. 1, the preview serves as "ground truth" of what frame 0 resembles and can substantially enhance the effectiveness of receiver error concealment. Similarly, even if frames -2, -1, and 0 are lost together with their previews, the reception of the preview for frame 1 is sufficient to reconstruct a likeness of frame 1, and can be immediately employed for concealment using the previously received frame -3. The preview layer is sometimes called the RECAP layer.

The RECAP layer plays a role not unlike auxiliary information for error resilient schemes using a distributed source coding (DSC) framework. SLEP, proposed by Rane *et al.* [2], provides a way to protect against isolated losses within a frame, but generally does not protect against burst loss of one or more frames. Wang *et al.* [3] proposed a scheme which attempts to mitigate error propagation, but its performance depends on the quality of the error concealed frame used as side-information in the error correction process and fails if correlation statistics are not accurately estimated. To utilize all available frames at the decoder, RECAP employs decoder motion search similar to PRISM [4], but is compatible with H.264 which has superior compression efficiency.

The rest of the paper is organized as follows. We describe one way an acknowledged preview can be used for receiver error concealment in Section 2. The effectiveness of RECAP against isolated loss and complete frame loss are evaluated in Section 3, followed by the conclusion.

2. A RECAP IMPLEMENTATION

The effectiveness of RECAP relies on three guiding principles. First, many lost blocks may appear in previous correctly reconstructed frames at the decoder. The decoder can exploit the available preview to reliably search for a suitable high-resolution block as replacement. Second, some missing blocks may be novel, *e.g.* when a previously occluded region is uncovered. The decoder can form a coarse reconstruction from the preview. Third, the preview should be low-rate and should not be affected by errors in previous frames. A low-rate overhead is achieved through low-resolution, while reliability is established via prediction on acknowledged frames only. A block diagram for the processes is shown in Fig. 2.

2.1. Encoder

The main presentation is encoded using H.264, whose output pictures are down-sampled by 4 in each dimension to yield preview pictures 16 times smaller for low rate and low encoding complexity overhead. These preview pictures are then encoded with another H.264 encoder but using only positively acknowledged pictures as reference frames for reliability.

2.2. Decoder

The RECAP decoder first decodes the received video bitstream to obtain a reconstruction of each frame. This could have corrupted regions or drift artifacts due to either current or prior transmission errors. The RECAP decoder also decodes the received RECAP bitstream to obtain the preview frame and sends an ACK feedback message.

After a transmission error occurs, the decoder compares the preview frame and the reconstructed frame to localize blocks which are not consistent with the preview and hence would appear visually "wrong". This is achieved by examining each non-overlapping 16×16 block of pixels in the reconstructed frame, and computing the mean square error (MSE) between its down-sampled 4×4 block of pixels and the co-located block in the preview frame, *e.g.* for block A in Fig. 3,

the MSE between block B and C is computed. A threshold is then applied to determine if the full-resolution block is consistent with the preview. The threshold is computed from past correctly received frames and preview to adapt to both preview reconstruction quality and video content.

In the RECAP error concealment step, one or more received frames that are error-free can be used as reference frames. Each reference frame is smoothed with the same anti-aliasing filter used in the down-sampler to avoid inadvertent inclusion of high spatial frequency during subsequent decoder motion search. When a block is determined to be inconsistent with the video preview, its preview block, *e.g.* block C in Fig. 3, is used as a descriptor in performing motion search on the smoothed reference frame. The predictor block with the smallest MSE, *e.g.* block D in Fig. 3, is one candidate for error concealment. If the smallest MSE is too high, then the up-sampled version of the preview block, *e.g.* block E in Fig. 3, will be selected for error concealment instead. The final error concealed frame is placed in the reference frame buffer of the full-resolution H.264 decoder for subsequent decoding.

3. EXPERIMENTAL RESULTS

In our experiments, we compare the proposed RECAP scheme with (i) H.264; (ii) H.264 with FMO using a checkerboard pattern; and (iii) H.264 with UEP on the macroblock mode and motion vectors (*i.e.* data partition A [5]). We used a 720p 30fps video sequence typical of a video conferencing scene divided into three portions, where a man walks in ("Enter"), sits down ("Sit"), and waves at the camera ("Wave"). The video is coded at a rate of about 3.25 Mbps, with 6 slices per frame that are mapped to 6 transport packets. The RECAP bitstream is transported in a separate packet, as is the parity data in the UEP scheme. We used a RTT equivalent to 7 frames of video, hence the preview stream typically use reference frames that is 7 frames ago. We target the RECAP bitstream to use 7% of the full-resolution video bit-rate. For this sequence, FMO uses an additional 5% rate, while UEP, with 2 parity packets, uses an additional 8% rate.

3.1. Single packet loss

We first consider the case where only a single packet is lost during transmission. Previous results suggest that both FMO and UEP would perform fairly well [5]. For RECAP, the reconstruction quality depends on whether the preview for the lost packet is also lost. Here, we consider a worst case scenario where the RECAP packet for the frame with error is also lost, *e.g.* frame 0' in Fig. 1 is not received.

Fig. 4 shows how PSNR varies with frame after the loss of a single packet, denoted by frame 0 in the plots, for various clips. In "Enter", because background is uncovered just as the man walks in, UEP does not do a good job of concealing that region. While FMO does a much better job of concealment, some errors still remain. Thus, for both of these methods, drift occurs and a large part of the background is corrupted, leading to the decreasing PSNR. On the other hand, RECAP is able to reconstruct much of the background, either by copying or up-sampling, leading to a improved PSNR in frame 1, and maintains that video quality. As discussed above, RECAP could have improved the PSNR in frame 0 as well by making use of the same scheme, but in these experiments we choose not to in order to consider the worst case scenario. In "Sit" and "Wave", RECAP demonstrates that it is competitive with FMO and UEP respectively. In Table 1, we show the average PSNR for the 7 frames after the packet loss[1]. While RECAP does not always have the best performance, *e.g.* in "Sit", the gap is not too large.

Table 1. Average PSNR for 7 frames following error for isolated packet loss. The best result is shown in bold, and difference in bracket.

Scheme	"Enter"	"Sit"	"Wave"
H.264	28.74 (-5.87)	30.74 (-2.35)	33.71 (-1.47)
FMO	32.53 (-2.08)	**33.09**	34.25 (-0.93)
UEP	29.48 (-5.13)	32.68 (-0.41)	**35.18**
RECAP	**34.61**	32.83 (-0.26)	**35.18**

Table 2. Average PSNR for 7 frames following error for complete frame loss. The best recovery is shown in bold, with difference shown in bracket. Note that UEP and H.264 will have the same performance.

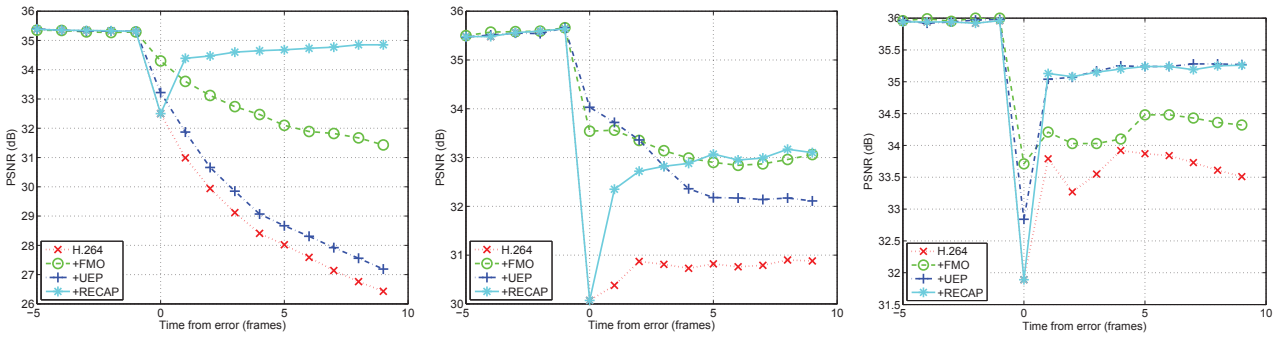
Scheme	"Enter"	"Sit"	"Wave"
H.264	26.62 (-6.59)	28.91 (-3.92)	31.94 (-2.40)
FMO	26.26 (-6.95)	28.66 (-4.17)	31.82 (-2.52)
RECAP	**33.21**	**32.83**	**34.34**

3.2. Complete frame loss

Next, we consider the case when a burst loss wipes out all slices in a frame. Fig. 5 shows how PSNR varies with frame after a complete frame loss, denoted by frame 0 in the plots, for various clips. Results for UEP are the same as plain H.264, since none of the macroblock mode or motion information can be retrieved, and is omitted. FMO offers little or no improvement over simple H.264 since no neighboring macroblock is available to aid error concealment. In contrast, RECAP excels in error concealment, as evidenced by the quick recovery in quality of frame 1 in each of the 3 cases, with no degradation of video quality in subsequent frames. Again, this is due to the fact that RECAP is able to either copy the right full-resolution block or replace with a coarsely reconstructed version of the block for error concealment. Table 2 presents the average PSNR for the 7 frames after the complete frame loss, showing clearly the large performance gap between RECAP and FMO.

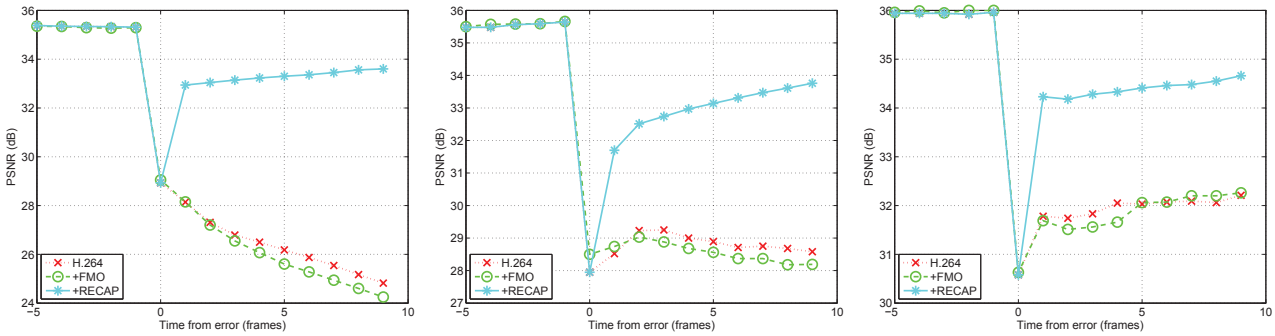
In Figs. 4 and 5, RECAP not only has generally higher PSNR, but the duration of when PSNR is low (say below 32 dB) is short. The decoder thus can hide the more corrupted

[1] If RPS is used, we would expect video quality to recover after one RTT.



(a) Enter (b) Sit (c) Wave

Fig. 4. **PSNR evolution for single packet loss.**



(a) Enter (b) Sit (c) Wave

Fig. 5. **PSNR evolution for complete frame loss.**

frames by simply repeating the last good frame without perceivable freeze. This works in a shorter time scale than the reaction time of RPS would permit and helps maintain break-up and freeze free streaming even when RTT is significant.

4. CONCLUSION

In this paper, we have presented the motivation and description for RECAP, a scheme that improves error resilience by sending an additional robustly coded low-rate and low-resolution description of the source video. In addition to providing a descriptor that can be used for decoder motion search or up-sampled to provide a coarse reconstruction, RECAP also enables generating a video preview reliably with low-complexity. We show promising results for RECAP which suggest that RECAP works well for both isolated packet loss and complete frame loss, unlike FMO and UEP which are only effective for isolated packet loss.

The implementation described here can be improved in several ways. First, it is likely that enforcing spatial coherence of motion, by using an over-lapped block search or otherwise, would improve performance. Second, both error detection and the decision between a motion-compensated predictor and up-sampled predictor could be approached in a more principled manner. For example, an estimation theoretic framework suggested for scalable video coding by Rose and Regunathan [6] could be applied to estimate the missing macroblock. Third, an additional Wyner-Ziv layer can be used to improve the quality of the error-concealed frame [7].

5. REFERENCES

[1] Yao Wang and Qin-Fan Zhu, "Error control and concealment for video communication: a review," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 974–997, May 1998.

[2] S. Rane, P. Baccichet, and B. Girod, "Modeling and Optimization of a Systematic Lossy Error Protection System based on H. 264/AVC Redundant Slices," in *Proc. Picture Coding Symposium (PCS)*, 2006.

[3] J. Wang, A. Majumdar, K. Ramchandran, and H. Garudadri, "Robust video transmission over a lossy network using a distributed source coded auxiliary channel," in *Proc. Picture Coding Symposium (PCS)*, 2004.

[4] R. Puri, A. Majumdar, and K. Ramchandran, "PRISM: A Video Coding Paradigm With Motion Estimation at the Decoder," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2436–2448, Oct 2007.

[5] S. Wenger, "H.264/AVC over IP," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 645–656, July 2003.

[6] K. Rose and SL Regunathan, "Toward optimality in scalable predictive coding," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 965–976, Jul 2001.

[7] D. Mukherjee, B. Macchiavello, and R.L. de Queiroz, "A simple reversed-complexity Wyner-Ziv video coding mode based on a spatial reduction framework," in *Proc. SPIE Visual Communications and Image Processing*, 2007, vol. 6508, p. 65.