ROBUST LOW-DELAY VIDEO TRANSMISSION USING H.264/AVC REDUNDANT SLICES AND FLEXIBLE MACROBLOCK ORDERING

Pierpaolo Baccichet[*], *Shantanu Rane*[†], *Antonio Chimienti*[‡] *and Bernd Girod*[†]

[*] Max Planck Center for Visual Computing & Communication, Stanford, CA 94305
[†] Information Systems Laboratory, Stanford University, Stanford, CA 94305
[‡] I.E.I.I.T - Consiglio Nazionale delle Ricerche, Torino, Italy
{bacci,srane,bgirod}@stanford.edu

ABSTRACT

This paper proposes a scheme for low-delay robust transmission of video signals over packet erasure channels. In applications such as video conferencing, the permissible delay between encoding and playback may be too low to allow retransmission or channel coding approaches which require buffering several video packets. For such a scenario, we present a scheme that provides error robustness using redundant video descriptions applied to pertinent portions of the video signal. In the H.264/AVC specification, this can be efficiently implemented using redundant slices and Flexible Macroblock Ordering (FMO). We describe a model that determines the bit rate of the redundant descriptions such that the expected distortion at the decoder is minimized. Across all the video test sequences used, the average video quality delivered by the proposed scheme is 3.7 dB higher than decoder-based error concealment, and 1.2 dB higher than encoder-based loss-aware rate-distortion optimization.

Index Terms— Error-resilient video coding, H.264/AVC, flexible macroblock ordering, redundant slices

1. MOTIVATION

This paper is concerned with low-delay robust video transmission. Applications which fall into this area include Internet conversational video services, and low-delay live video streaming over wireless or wired networks. Because of the low-delay requirement, it may not always be permissible to rely on retransmissions to recover from packet losses. Most modern decoders are capable of performing error concealment to mitigate losses. However, owing to limited available information and limited signal processing resources, decoder-based error concealment does not generally provide acceptable video quality.

Feedback can be used to notify the encoder of the losses, and to adapt the mode decisions, in particular the selection of the reference picture, within a loss-aware rate-distortion-optimal framework [1]. This technique incurs delays of less than 1 second, making it suitable for video-conferencing. Alternatively, one or more frames may be protected using systematic source/channel coding approaches such as FEC [2]. In our own work [3, 4], Systematic Lossy Error Protection (SLEP) has been shown to provide a more flexible resilience-quality tradeoff and more graceful degradation than conventional FEC. SLEP is based on applying a channel code to coarsely quantized redundant video descriptions. However, when the bit rate of the video signal is low, several frames must be buffered before channel encoding/decoding can be performed, resulting in a large delay between encoding and playback. The scheme presented in this paper also uses redundant video descriptions, but avoids channel encoding/decoding to provide robustness with very low delay.

In this work, we describe a scheme in which coarsely quantized redundant descriptions are generated for certain portions of the video signal, and are used to provide error resilience when the primary video signal is lost. An implementation using H.264/AVC is proposed, which leverages standard-compliant tools, namely, redundant slices and Flexible Macroblock Ordering (FMO). Using a model for the average end-to-end video quality, the bit rate used to encode the redundant descriptions is optimized. Further, FMO enables the encoder to choose the way in which this bit rate is allocated to the redundant description, e.g., encoding either a redundant description of the entire video frame or only a region-of-interest within the frame.

The remainder of this paper is organized as follows: The encoding and decoding scheme for the redundant descriptions is described in Section 2. Section 3 explains how to choose the encoding parameters for the redundant description. In Section 4, the robustness of the proposed scheme is investigated experimentally.

2. ERROR RESILIENT VIDEO CODING SCHEME

An H.264/AVC implementation of the proposed error resilience scheme is shown in Fig. 1. The steps used to encode the redundant description are as follows:

Determine Redundant Slice Map: This step involves determining which portions of the video frame will be encoded using redundant slices. In the current implementation, one of three options may be chosen:

1. Encode the entire frame as a succession of redundant slices. This is a trivial map in which there are one or more redundant slices which are, in turn, placed into one slice group.[1]
2. Identify a Region-Of-Interest (ROI) at the encoder, as shown in Fig. 2. The ROI is specified in the bit stream using FMO Type 2. This involves covering up the non-ROI region with up to 7 slice groups. The remaining portion of the video frame now consists of the ROI. Only this latter portion is encoded into redundant slices. A detailed description of the method used to determine the ROI and to specify it in the bit stream using FMO Type 2, can be found in [5].
3. Define two slice groups, one containing even rows of macroblocks, and the other containing the odd rows. Redundant slices are generated for both slice groups.

The Redundant Slice Map has to be specified per frame of the video sequence, and travels inside a standardized container known as the Picture Parameter Set (PPS). The method by which one of the above choices is preferred over the other two, is specified in the next section. Typically, it is observed that the ROI map is selected for sequences with a static background (e.g., Akiyo), the Even-Odd map is selected for sequences in which there is high motion in a dominant direction (e.g., Bus), while the trivial redundant map in choice 1 above is selected in most other cases.

[1] In the H.264/AVC nomenclature, a frame is composed of one or more slice groups, and a slice group is composed of one or more slices.

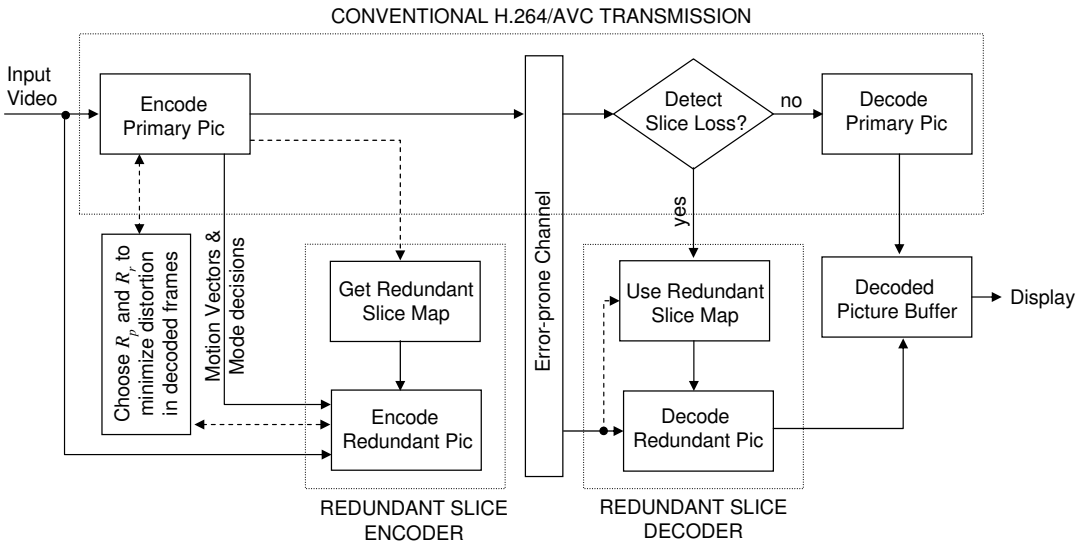


Fig. 1. Within the H.264/AVC framework, a portion of the video signal is specified using Flexible Macroblock Ordering (FMO) and redundantly encoded using coarse quantization. When slices from the primary coded picture are lost, the redundant slices limit the maximum degradation that can occur. A model selects the encoding bit rates of the primary and redundant descriptions, such that the average distortion in the decoded frames is minimized.

Perform Redundant Slice Encoding: This step takes the redundant slice map as the input, and encodes the region specified by it, into one or more redundant slices. Note that the redundant slices are of a fixed maximum length specified at the encoder. Thus, they have arbitrary shapes (independent of the primary slices) depending upon the complexity of the scene, the coding modes, and the target bit rate. To reduce complexity, we impose a constraint that the macroblocks in the redundant slices must have the same coding modes and motion vectors as those used in the primary slices. This same strategy was used in [3]. The quantization step size used to encode the redundant slices is constrained to be higher than that used in the primary slices. This reduces the bit rate of the redundant slices, at the expense of a small quantization mismatch. When primary slices are lost, the decoder attempts to conceal the losses using the redundant description. If the redundant slices are also lost, then there is no choice but to use some decoder-based error concealment scheme. In the current implementation, the non-normative concealment algorithm in H.264/AVC [6] is used for this purpose.

3. MODELING AND OPTIMIZATION

We now describe a model for the average end-to-end distortion incurred by using H.264/AVC redundant slices. The model is adopted from our earlier work [3, 7], in which the derivation is provided in detail and the accuracy of the model is established. Let the distortion-rate pairs for encoding the primary and redundant pictures be denoted by (D_p, R_p) and (D_r, R_r) respectively. As the redundant slices are coarsely quantized compared to the primary slices, $R_r \leq R_p$, $D_r \geq D_p$.

Let $D[i]$ be the *average* end-to-end MSE experienced by a packet in the i^{th} frame (assume it is a P frame). We consider three distinct scenarios: (1) There are no errors, and error energy in frame i is contributed only by the distortion propagating from the previous frame, denoted as $D[i-1]$, (2) The primary slice is lost, but is concealed using its corresponding redundant slice. The total distortion contribution from error propagation and redundant slice concealment is $D[i-1] + D_r - D_p$, with $D_r - D_p$ representing the error energy corresponding to the quantization mismatch between the primary and redundant descriptions, (3) Both the primary and redundant slices are lost. The resulting distortion from error propagation and previous frame error concealment is modeled as $D[i-1] + \text{MSE}[i, i-1]$, where $\text{MSE}[i, i-1]$ is the mean squared error between frames i and $i-1$. The derivation of these three distortions by averaging per-pixel squared errors is explained in detail in [7]. Combining the three distortions and weighting each by its probability of occurrence,

$$\begin{aligned} D[i] = {} & (1-p)\, D[i-1] + p\,(1-p)\,(D[i-1] + D_r - D_p) \\ & + p^2\,(D[i-1] + \text{MSE}[i, i-1]) \end{aligned} \tag{1}$$

where p is the packet erasure probability seen by the video decoder at the application layer. This relation clearly demonstrates the effect of error propagation, which can be mitigated by insertion of intra macroblocks. If a macroblock is refreshed every N frames, then the average MSE over N frames is

$$\mathcal{D} = \frac{1}{N} \sum_{i=1}^{N} D[i] \tag{2}$$

The objective is to select R_p and R_r, the encoding bit rates of the primary and the redundant descriptions, such that $\mathcal{D}$ is minimized. To do this, it is necessary to have a model for the distortion-rate functions $D_p(R_p)$ and $D_r(R_r)$. We use the model proposed in [8], in which the two distortion-rate functions can be modeled using three parameters each, as follows:

$$D_p = D_{0p} + \frac{\theta_p}{R_p - R_{0p}} \;,\; D_r = D_{0r} + \frac{\theta_r}{R_r - R_{0r}} \tag{3}$$

The parameters, θ_p, θ_r, R_{0p}, R_{0r}, D_{0p}, and D_{0r} are determined from trial encodings. For a total bit rate constraint R_T, the encoder solves the following optimization problem:

$$\text{Minimize } \mathcal{D}(R_p, R_r) \;\; \text{such that } R_p + R_r \leq R_T \tag{4}$$

where $\mathcal{D}$ is the average end-to-end MSE given by (2).

It would be sufficient to use the above source coding model for $D_r(R_r)$, if there were only one method for encoding the redundant

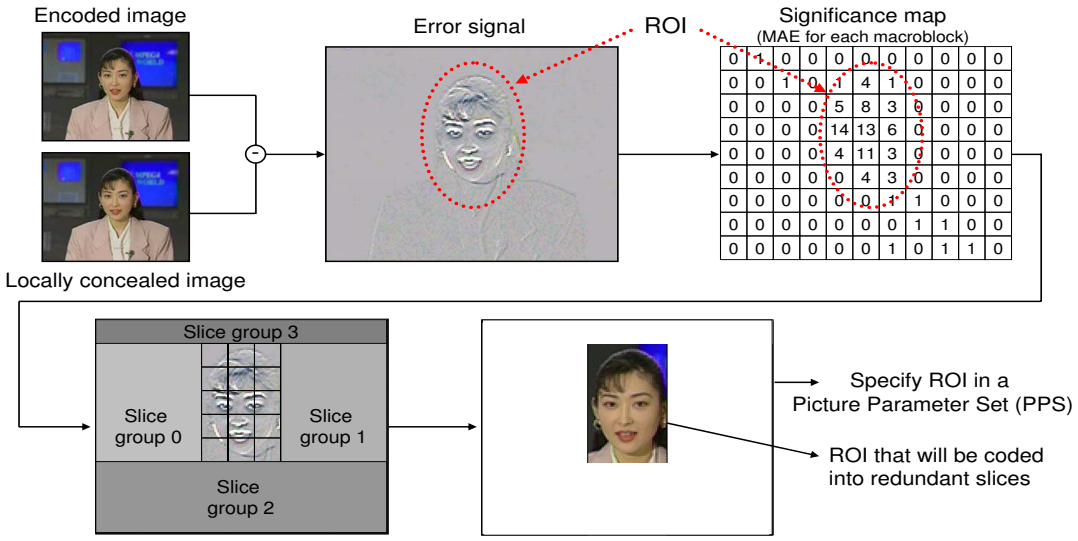


Fig. 2. A region-of-interest (ROI) is determined by performing simple error concealment at the video encoder, and then choosing the macroblocks which are the largest contributors to the concealment errors. This ROI is then encoded into redundant slices. The mapping of an image area into ROI and non-ROI can be efficiently specified in H.264/AVC using FMO type 2 (Foreground With Leftover)

Sequence	Total Bit Rate(kb/s)
Bus.CIF	1000
Foreman.CIF	500
Akiyo.CIF	200
Mother Daughter.CIF	300

Table 1. Video sequences used in the simulations. The maximum allowable size of a video slice is set to 500 bytes, and each sequence is 1000 frames long and encoded at 15 frames/s.

slices. However, as explained in Section 2, there are three possible ways to encode the redundant slices, and potentially there could be other redundant slice maps. It would be prohibitively complex to determine three encoder parameters θ_r, R_{0r} and D_{0r} for each redundant slice map. Therefore, we make a compromise as follows: The parameters are determined only for the trivial slice map, in which the entire frame is considered as a slice group. The optimization in (4) is carried out and optimum bit rates R_p^* and R_r^* are obtained. Then, a trial encoding is performed for each non-trivial redundant slice map. The redundant slice map which gives the smallest MSE for bit rate R_r^*, is then adopted for the current frame.

A further reduction in complexity is obtained by performing the optimization once every N frames, where N is the intra refresh period for a macroblock. More precisely, R_p and R_r are optimized once every N frames, but the redundant slice map is updated for every frame.

4. EXPERIMENTAL RESULTS

The robustness of the proposed scheme was investigated for a channel that introduces random symbol errors (1 symbol = 1 byte). We compared the average and instantaneous video quality delivered to the decoder by (a) The proposed scheme which adaptively changes the redundant slice map, (b) A scheme which optimizes the redundant bit rate but always chooses to redundantly encode the full frame, (c) Decoder-based error concealment [6], and (d) Encoder-based Loss-Aware Rate-Distortion Optimization (LA-RDO) [9].

All the encoder-based schemes were optimized for a symbol error probability of 10^{-4}, and the performance was tested over a range of error probabilities from 2×10^{-5} to 5×10^{-4}. To mitigate the effect of instantaneous variations in the statistics of the error process,

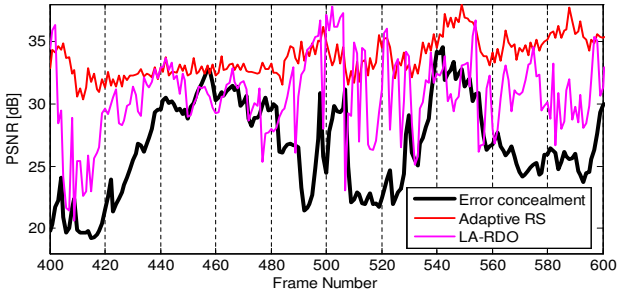
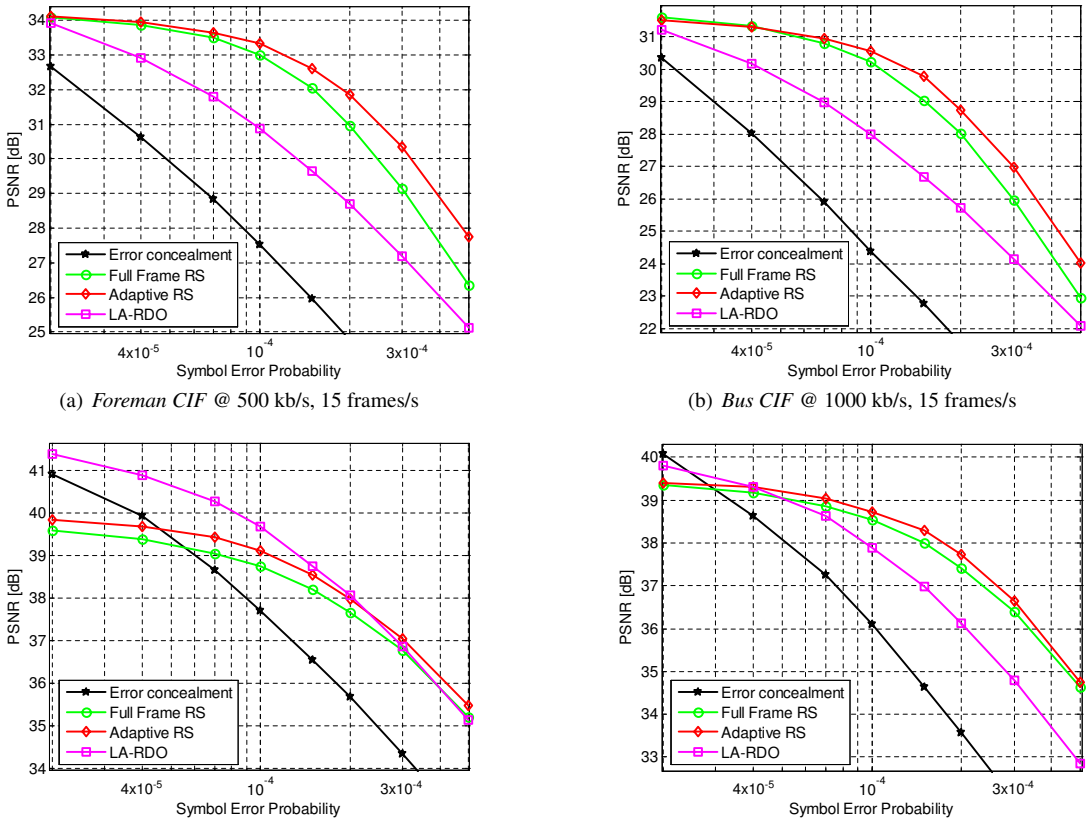


Fig. 4. Redundant slices do not incur the high distortion associated with error concealment artifacts, as seen here for the *Foreman* sequence. An adaptive redundant slice map gives better instantaneous PSNR compared to trivially including the entire frame inside one redundant slice group.

the received picture quality was averaged over 10 realizations of the channel. The sequences used in the experiments are listed in Table 1. Since introducing an intra frame would cause a sudden increase in the bit rate, it was decided to perform intra refresh of one row of macroblocks per frame. This amounts to a full intra refresh every 18 frames of a CIF sequence.

The average PSNR of the decoded video sequence is plotted against the symbol error rate in Fig. 3. The proposed scheme provides significantly higher PSNR than both decoder based error concealment and LA-RDO. Further, adaptively changing the redundant slice map improves the average picture quality by up to 0.4 dB for sequences with low motion, such as *Akiyo* and *Mother Daughter*, and by up to 1.4 dB for *Foreman* and *Bus*, which have higher motion. To accommodate the redundant slices, the primary bit rate must be reduced, resulting in a loss in video quality at low symbol error probabilities, as seen in Fig. 3(c). This reduction in picture quality is dependent on the bit rate chosen by the model for encoding the redundant slices, i.e., on the quantization mismatch between the primary and redundant slices. The computational complexity required to encode the redundant slices is significantly lower than that required to implement LA-RDO, which performs several channel simulations and decodings at the encoder. Across all four video sequences, and the entire range of symbol error probabilities, the average video quality of the proposed



(a) *Foreman CIF* @ 500 kb/s, 15 frames/s

(b) *Bus CIF* @ 1000 kb/s, 15 frames/s

(c) *Akiyo CIF* @ 200 kb/s, 15 frames/s

(d) *Mother Daughter CIF* @ 300 kb/s, 15 frames/s

Fig. 3. The average received PSNR degrades gracefully when the symbol error probability increases. Moreover, choosing the redundant slice map adaptively results in improved picture quality as compared to trivially encoding the entire frame as one redundant slice group. For high motion sequences (*Bus* and *Foreman*), and the scheme favors the trivial (full-frame) slice map or the even-odd slice map. For low motion sequences (*Akiyo* and *Mother Daughter*), the scheme favors the ROI slice map.

scheme is 3.7 dB higher than decoder-based error concealment, and 1.2 dB higher than LA-RDO.

Finally, Fig. 4 plots the luminance PSNR versus frame number for the *Foreman* sequence at a symbol error probability of 10^{-4}. It is observed that, in comparison with LA-RDO and error concealment, redundant slices are able to mitigate the instantaneous reduction in the frame PSNR when the primary slices are lost.

5. SUMMARY

Redundant video descriptions can be used to design a robust video coding scheme with a very low encoding/decoding delay. An H.264/AVC implementation using redundant slices and Flexible Macroblock Ordering (FMO) has been proposed. In this scheme, a coarsely quantized redundant description is generated for certain portions of a video frame. When primary slices are lost, the corresponding redundant slices are used to conceal the losses up to a certain residual distortion, which depends on the quantization mismatch between primary and redundant slices. The available rate is allocated to the primary and redundant descriptions so as to minimize the average distortion in the received video. Out of three different redundant slice mappings, the encoder selects the one with the minimum MSE reconstruction for the given rate allocation. Experimental results demonstrate that the proposed scheme provides higher instantaneous and average picture quality than both decoder-based error concealment and encoder-based loss-aware rate-distortion optimization.

6. REFERENCES

[1] Y. Liang and B. Girod, "Network-Adaptive Low-Latency Video Communication over Best-Effort Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 72–81, Jan. 2006.

[2] Y. Wang and Q. Zhu, "Error Control and Concealment for Video Communications: A Review," in *Proceedings of the IEEE special issue on Multimedia Signal Processing*, May 1998, vol. 86, pp. 974–977.

[3] S. Rane, P. Baccichet, and B. Girod, "Modeling and Optimization of a Systematic Lossy Error Protection System," in *Picture Coding Symposium (PCS 2006)*, Beijing, China, Apr. 2006.

[4] P. Baccichet, S. Rane, and B. Girod, "Systematic Lossy Error Protection using H.264/AVC Redundant Slices and Flexible Macroblock Ordering," in *Proc. IEEE Packet Video Workshop*, Hangzhou, China, Apr. 2006.

[5] P. Baccichet, *Solutions for the Protection and Reconstruction of H.264/AVC Video Signals for Real-Time Transmission over Lossy Networks*, Ph.D. dissertation, University of Milan, Milan, Italy, 2006.

[6] Y. Wang, M. Hannuksela, V. Varsa, A. Hourunranta, and M. Gabbouj, "The Error Concealment Feature in the H.26L Test Model," in *Proc. IEEE International Conference on Image Processing*, Rochester, NY, Sept. 2002, vol. 2, pp. 729–732.

[7] S. Rane and B. Girod, "Analysis of Error-Resilient Video Transmission based on Systematic Source-Channel Coding," in *Picture Coding Symposium (PCS 2004)*, San Francisco, CA, Dec. 2004.

[8] K. Stuhlmüller, N. Färber, M. Link, and B. Girod, "Analysis of Video Transmission over Lossy Channels," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1012–1032, June 2000.

[9] T. Stockhammer, D. Kontopodis, and T. Wiegand, "Rate-Distortion for JVT/H.26L Video Coding in Packet Loss Environment," in *Proceedings of the 2002 Packet Video Workshop, Pittsburgh, PA, U.S.A*, Apr. 2002.