MULTI-RESOLUTION REDUNDANCY FOR ERROR-RESILIENT VIDEO TRANSMISSION

Debargha Mukherjee and Wai-Tian Tan
HP Laboratories, Palo Alto, CA, USA.
Email: {debargha.mukherjee, wai-tian.tan}@hp.com

ABSTRACT

In this paper we advocate a multi-resolution mechanism for redundant information generation and transmission with low overheads in bit-rate, in order to enable reliable video communication over challenging lossy networks with low latency. Our previous work entitled RECAP, transmitted a low resolution redundant version of a parent video stream, in order to achieve effective concealment of isolated and burst losses by clever multi-frame super-resolution processing at the decoder. A reference picture selection mechanism was used to transmit the low-resolution redundant video with high guarantee and stop drift in case of losses. In this work, we extend the framework to incorporate an additional distributed Wyner-Ziv coding layer on the low-resolution information to further correct errors, and consequently improve the error-concealed picture quality. The error-concealed super-resolved frame using the low-resolution information alone, now acts as side-information at the decoder to correct additional errors by decoding the Wyner-Ziv layer. Preliminary results are presented to demonstrate the efficacy of the proposed approach.

Index Terms – Wyner-Ziv coding, source-channel coding, error-resilient video, multi-frame processing, error concealment.

1. INTRODUCTION

Conventional video coding schemes typically use predictive coding based on a set of previously encoded reference frames using motion compensation. Lossy unreliable networks pose enormous challenges in reliable transmission of such video [1]. The primary difficulty arises from predictive mismatch between a reference frame at the encoder and a corresponding reference frame at the decoder due to packet losses, resulting in drift and objectionable artifacts that persist until the next refresh achieved by either INTRA frames, SP frames or reactive Reference Picture Selection (RPS) [3]. While automatic repeat-request (ARQ) and forward error-correction schemes (FEC) [1] could be used to alleviate the problem, these schemes result in either inordinate delay (ARQ) or excessive redundant bit-rate due to the need to design a system for the worst-case (FEC). Schemes that partition the compressed data into critical and non-critical parts and protect the former more heavily, are less wasteful of bandwidth but are less effective in making artifacts unnoticeable.

One alternative to sending redundant FEC information along with the compressed bit-stream is to send a different kind of redundant information that does not yield the compressed bits directly after decoding, but rather aids the concealment operation performed at the decoder when packet losses are encountered. Our previously proposed framework called RECAP (Receiver Error-Concealment using Acknowledged Preview) [4] accomplished exactly that by sending a low resolution version of the original compressed video as redundant information. The redundant bit-stream is communicated using a high-guarantee drift-free mechanism, for example RPS. When losses are encountered in the parent stream, the low-resolution thumbnail is used as a low quality reference to search previous correctly decoded full-resolution frames in order to find the best match to replace missing/corrupted blocks. Such a form of super-resolving concealment works amazingly well in most cases, and the resultant drift is usually small enough for the artifacts to remain unobjectionable until the next INTRA refresh or reactive RPS to arrest error propagation. However, occasional glitches in certain parts of the frame where the concealment operation fails due to complex motion or fine details, are common. In this paper, we generalize the framework to further arrest these errors.

Inspired by the foundation of Slepian-Wolf [5] and Wyner-Ziv (WZ) [6] theorems, immense attention has been devoted to practical distributed source coding problems in recent years [7]-[18]. One area where such methods have been used is error-resilient video communication [10]-[13]. Here the main advantage is that decoding can be conducted reliably even when the reference frame is inexact, provided statistical relationships between inexact and exact references are known. Sehgal [10] proposed a scheme where periodic distributed coded peg frames are transmitted with the bit-stream to arrest drift due to loss in the network and obtain an exact reconstruction. Recently Rane [12][13] proposed a class of Systematic Lossy Error Protection (SLEP) schemes, where distributed source coding is used to recover a lower quality version of a frame when there is loss rather than arrest drift completely, resulting in reduced bit-rate overheads.

In order to improve the RECAP framework [4] by only limited additional bit-rate we adopt an approach similar to the SLEP scheme above. In other words, we attempt to only recover an approximate version of the corrupted frame, which nonetheless would be closer to the original than the one obtained by super-resolution processing alone. We add an additional distributed (Wyner-Ziv) coded layer to the low-resolution RECAP layer, which subsequently is decoded using the super-resolved error-concealed frame as side-information. The next section provides more details on the framework.

2. MULTI-RESOLUTION FRAMEWORK

We start with a brief review of the RECAP framework [4]. Given a compressed bit-stream, RECAP generates a redundant bit-stream by encoding a low resolution (LR) version of the reconstructed video with a standard encoder. Typically, the low resolution is limited to ¼ in each dimension, resulting in a redundant bit-stream of less that 10% of the bit-rate of the original bit-stream. The LR redundant bit-stream is communicated to the receiver using a high-guarantee drift-free mechanism. In the live point-to-point streaming case of [4], the high-guarantee mechanism for the LR layer used was pro-active RPS, where all frames are predicted only based on positively acknowledged frames by the receiver. By avoiding prediction using unacknowledged frames, every received frame can be decoded drift-free. It is conceivable however that other mechanisms could be used just as well.

When a frame or part thereof is lost due to losses in the network, the corresponding LR frame is used to conceal the error in locations corresponding to corrupted slices as follows. The LR reconstruction is used as a reference to blockwise search previous correctly decoded high-resolution (HR) frames after low pass filtering; if the match is good, the high-frequency component of the best matching HR block from previous frames is added to the spatially upsampled version of the LR block; if not, spatially upsampling the LR block is the better choice. This operation of super-resolving a LR frame using a set of high-resolution temporal neighbors as reference frames, has been termed semi-super-resolution [14] elsewhere, and has been explored more rigorously in [15][16][19]. For the purpose of this paper, we use a functional representation of this operation as follows:

978-1-4244-4296-6/10/$25.00 ©2010 IEEE 2326 ICASSP 2010
Fig. 1 Multi-resolution redundancy generation and decoding framework with (a) WZ layer operating at full-resolution; and (b) WZ layer operating at intermediate resolution. In both cases, \( X \) represents a reconstructed frame of a parent bit-stream not shown, and \( \tilde{X} \) represents the reconstruction.

\[ \tilde{X} = \Psi^{(e)}(X^{(i)}, \{R_e, R_i, \ldots, R_{e-1}\}) \],

where output \( \tilde{X} \) is a full-resolution super-resolved frame (block) obtained by operator \( \Psi^{(e)}(\cdot) \) operating on \( X^{(i)} \), which is a downsampled (and compressed) version of the full-resolution frame (block) by a factor \( r \) in each dimension, and a set of \( K \) full-resolution reference frames \( \{R_e, R_i, \ldots, R_{e-1}\} \) previously decoded correctly.

While the above LR-assisted error concealment strategy is generally quite effective, in certain situations involving complex motion or fine detail lines, it may not be possible to recover a visually faithful representation of the lost high-resolution image. In areas of unsuccessful match during decoder-motion search, the error concealment mechanism must rely on spatial interpolation from the low resolution preview layer to create the final error concealed version. Accordingly, higher frequencies of the image lost during encoding of the LR layer remain unrecovered. However, if we can afford to spend an additional 7-8\% of the bit-rate to convey some extra information, some of this lost information may be recovered.

The primary contribution of this work is incorporating an additional second layer of information over the baseline LR layer to achieve this goal. Since the LR layer already conveys the low-resolution content of the current frame, the second layer only needs to encode the Laplacian residue as a spatial enhancement layer. However, because the semi-super-resolution operation \( \Psi^{(e)}(\cdot) \) at the decoder would already have recovered most of the high-frequency information, we can save substantial bit-rate by transmitting only limited information corresponding to information that cannot be recovered. Unfortunately, the result of the \( \Psi^{(e)}(\cdot) \) operation depends on the particular loss pattern and the set of previous correctly decoded reference frames available at the decoder. Since this information is not deterministically available at the encoder, we need to use source coding with side-information (a form of distributed source coding) or Wyner-Ziv (WZ) coding techniques to code the Laplacian residue, using only statistical information. This layer is termed the WZ layer. The decoder recovers a corrupted/lost frame using the super-resolving operation \( \Psi^{(e)}(\cdot) \) on the LR layer, and then uses the residue computed from the resultant frame as side-information to decode the second layer. If \( h_{LR} \) and \( h_{WZ} \) respectively denote the LR and WZ layer bit-stream components of the redundancy respectively, the encoding and decoding operation can be depicted as in Fig. 1(a), where the LR layer is coded at resolution \( 2^d \) down from the full-resolution in each dimension.

In practice, coding the WZ layer at full resolution may be too wasteful of bandwidth. An alternative is to code the WZ layer at an intermediate resolution higher than the base layer, but less than full. Fig. 1(b) depicts this architecture, where the LR layer is coded at resolution \( 2^d \) down from full as before, but the WZ layer is coded at resolution \( 2^d \) down from full, with \( e < d \). In this architecture, the super-resolved frame is downsampled by a factor of \( 2^e \) before WZ decoding to obtain a reconstructed frame also at resolution \( 2^d \) down. A further super-resolving step is then needed to obtain the final full-resolution reconstruction.

In both architectures, the decoding operation can be iterated multiple times as follows. The final reconstruction \( \hat{X} \) can be downsampling by a factor of \( 2^e \) in each dimension and fed back into the decoding loop, replacing the base layer reconstruction \( \hat{X}^{(i)} \). Usually such an iterative process improves the decoding gradually over a few iterations without any additional bit-rate.

### 3. CODING OF THE WYNER-ZIV LAYER

Our Wyner-Ziv coder of the Laplacian residue frame operates in the transform domain, using an approach similar to that used in [18]. In other words, 8×8 DCT transform coefficients \( X \) are quantized by a uniform deadzone quantizer \( QP \) to obtain quantized values \( Q \), followed by a bi-level cost computation yielding an \( M \)-ary lower significant symbol \( C \) and an optional higher significant bit \( B \).

\[
\begin{align*}
Q &= \Phi(X, QP) = \text{sign}(X) \times \left\lfloor \frac{|X|}{QP} \right\rfloor, \\
\hat{X} &= \psi(Q, M) = Q - M \left\lfloor \frac{Q}{M} \right\rfloor + u(2Q - 2M) \left\lfloor \frac{Q}{M} \right\rfloor \quad (2) \\
& \quad \text{[Note : } u(k) = 1 \text{ for } k \geq 0; \text{ and } u(0) = 0 \text{ for } k < 0] \\
B &= \left\lceil \frac{Q}{M} \right\rfloor \mod 2
\end{align*}
\]

The \( M \)-ary least significant symbol \( C \) for all coefficients are source coded blockwise using a block coset entropy coder [18], while the most significant bits \( B \) are channel coded at a suitable rate \( R \) for each frequency by sending only the parity bits. Note that such a coding model is a special case of the case studied in [17], where one channel coded bit-plane is used at most.

The question that needs to be addressed for the above encoding model is how to choose the right parameters \( QP, M, R \) for a given target quality. In [17] we presented a detail analysis of parameter choice, encoding and decoding mechanisms for Wyner-Ziv coding of Laplacian sources under an additive noise model \( Y = X + Z \), where \( X \) is a Laplacian source with known variance \( \sigma_x^2 \), \( Z \) is an additive \( i.i.d. \) noise (Gaussian or Laplacian) independent of \( X \) with known variance \( \sigma_z^2 \), and \( Y \) is the side-information available at the decoder. Since DCT coefficients of the Laplacian residue are roughly Laplacian, [17] readily apply to this problem. For the RECAP-WZ problem, we find that the block DCT AC coefficients \( X \) and the corresponding coefficients \( Y \) from the super-resolved image, are in fact related by:

\[
Y = \rho X + Z, \quad \text{where } 0 < \rho < 1 \text{ is an attenuation factor that decays fast for higher frequencies, and } Z \text{ is a noise term which is closer to Laplacian in distribution.}
\]

This model is intuitive since the super-resolving operation not only loses some high-frequency information, but also...
adds noise in failure areas. Rewriting the model as \( Y/p = X + Z/p \), the same procedure [17] can be used to obtain the \( QP, M, R \) parameters, with \( (\sigma/p)^2 \) replacing \( \sigma^2 \) and \( Y/p \) replacing \( Y \) by during decoding.

The other issue to consider is computing the values of \( \rho, \sigma_X, \sigma_Y \). Fortunately, the std. dev. \( \sigma_Y \) for each frequency can be computed for the frame to be coded and transmitted as part of the WZ layer bit-stream. For the attenuation factor \( \rho \) and noise std. dev. \( \sigma_Y \), we train using \( \{X, Y\} \) pairs for each frequency and color component taken from a set of training videos. \( \sigma_Y \) is modeled as proportional to \( \sigma_X \), where the training process yields the proportionality factor.

The flow of parameter choice followed by encoding is depicted in Fig. 2 [18]. The parameter selection mechanism yields the \( QP, M, R \)-matrices, based on which the cosets \( C \) and bit-plane \( B \) are computed. The cosets \( C \) are block coset entropy coded [18]; \( B \) is aggregated over the frame for all frequencies and channel coded at average rate \( \hat{R} \) using an LDPC code. The parameter selection process is repeated at the decoder. Based on that cosets \( C \) are entropy decoded, and then bits \( B \) are soft decoded given side-information \( Y \) and \( Y \). Optimal reconstruction based on \( Y \) yields the final reconstruction for a coefficient.

In order to allow finer adaptation of the coding process from block to block, each 16x16 coding macro-block of the Laplacian residue is classified into one of several classes based on a combination of two features: (1) Sum of energies of the AC coefficients in it, and (2) Sum of energies of the AC coefficients in the co-located macro-block of the spatially upsampled LR layer reconstruction. It turns out that the super-resolving operation is most likely to fail if the AC energy in the source block is high, but the corresponding energy in the LR layer is low. Thus a ratio of the two is a good indicator of whether and how much additional information needs to be sent for a macroblock. Based on this ratio, macroblocks are classified into a few classes, and the class index is transmitted for each macroblock. Off-line training for the statistics are conducted for each class and the parameters are stored in the encoder and the decoder. For every WZ layer coded the source std. devs. \( \sigma_X \) are transmitted for each class, which along with offline trained parameters yield the coding parameters for each class matched to its statistics. Thus every macroblock is coded differently based on its classification.

We find that typically, only a few of the macroblocks in a frame actually gain from additional information. Consistent with that, our lowest class most often chooses zero-rate coding for a macroblock, and this class also has the highest population within a frame.

4. RESULTS

We tested our framework on standard 720p test sequences Mobile_Calendar and Shields coded using H.264 High Profile under isolated complete frame losses. For both cases, we used the architecture of Fig. 1(b). For the Mobile_Calendar sequence our LR layer was coded at \( \frac{1}{4} \) full 720p resolution in each dimension for a total bit-rate of about 10% of the parent bit-stream using a parallel H.264 encoder. We simulated a frame loss at frame 101 in the parent stream, and attempted to recover the frame using the LR layer, which is assumed to be transmitted error-free. The WZ layer is coded at \( \frac{1}{2} \) the 720p resolution, and uses only about 7% of the parent bit-stream, so that the overall redundancy is only 17%. Table 1 shows the recovery PSNR achieved by the LR and WZ layers over a few decoding iterations. The PSNR values shown are computed from the average MSE of \( Y, Cb \) and \( Cr \) components. The PSNR (\( \frac{1}{2} \) res) column in the table corresponds to the PSNR computed between the \( \frac{1}{2} \) resolution WZ reconstruction \( \hat{X}_{\frac{1}{2}} \) and \( X'_{\frac{1}{2}} \) in Fig. 1(b). The PSNR (full-res) corresponds to the PSNR computed between \( \hat{X} \) after super-resolving and \( X \). Table 2 shows the corresponding results for the Shields sequence for recovery of lost Frame 66, using LR layer redundancy of 10% and WZ layer redundancy of 8%.

Table 1. PSNR for recovery of Mobile_Calendar Frame 101, compared against parent bit-stream reconstruction, using redundant LR (10%) and WZ (8%) layers.

<table>
<thead>
<tr>
<th>Frame</th>
<th>LR Layer only</th>
<th>LR+ZL Layer – Iter 1</th>
<th>LR+ZL Layer – Iter 2</th>
<th>LR+ZL Layer – Iter 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR (( \frac{1}{2} ) res)</td>
<td>34.831</td>
<td>34.948</td>
<td>35.143</td>
<td>35.236</td>
</tr>
<tr>
<td>PSNR (full-res)</td>
<td>32.023</td>
<td>32.278</td>
<td>32.373</td>
<td>32.369</td>
</tr>
</tbody>
</table>

Table 2. PSNR for recovery of Shields Frame 66, compared against parent bit-stream reconstruction, using redundant LR (10%) and WZ (8%) layers.

<table>
<thead>
<tr>
<th>Frame</th>
<th>LR Layer only</th>
<th>LR+ZL Layer – Iter 1</th>
<th>LR+ZL Layer – Iter 2</th>
<th>LR+ZL Layer – Iter 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR (( \frac{1}{2} ) res)</td>
<td>34.081</td>
<td>34.290</td>
<td>34.772</td>
<td>35.236</td>
</tr>
<tr>
<td>PSNR (full-res)</td>
<td>31.916</td>
<td>32.408</td>
<td>32.436</td>
<td>32.340</td>
</tr>
</tbody>
</table>

In both cases we observe an anomaly that after the 3rd iteration PSNR at full-resolution actually drops even though the corresponding PSNR at 1/2 resolution still improves, which probably is an anomaly of the super-resolving mechanism used.

Fig. 4 shows a blow-up of the Mobile_Calendar sequence where the addition of the WZ layer is able to recover some of the fine detail lost by frame loss and subsequent super-resolving operation.

Fig. 3 shows the PSNR characteristics after a complete frame loss in the parent bit-stream, for the following two cases: (a) when the corresponding redundant data is available and used to recover the loss; (b) when the redundant data corresponding to the lost frame is lost as well, but drift in the LR part is arrested at the very next frame by use of RPS, and is subsequently used to recover the corrupt full-resolution frames. In both cases, the WZ layer is decoded only for the first lost frame for which the corresponding redundant information is available.

5. CONCLUSION

We have presented a framework for redundant information.
generation in multiple resolutions, where the base layer is pure source coded and the higher layer is distributed coded. Preliminary results are presented to show the benefits of the approach. We plan to continue explorations on the general framework in the future.

6. REFERENCES


