Scalable Video Coding with Robust Mode Selection

Rui Zhang, Shankar L. Regunathan and Kenneth Rose
Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106

Abstract

We propose to improve the packet loss resilience of scalable video coding. An algorithm for optimal coding mode selection for the base and enhancement layers is developed, which limits error propagation due to packet loss, while retaining compression efficiency. We first derive a method to estimate the overall decoder distortion, which includes the effects of quantization, packet loss and error concealment employed at the decoder. The estimate accounts for temporal and spatial error propagation due to motion compensated prediction, and computes the expected distortion precisely *per pixel*. The distortion estimate is incorporated within a rate-distortion framework to optimally select the coding mode as well as quantization step size for the macroblocks in each layer. Simulation results show substantial performance gains for both base and enhancement layers.

I. Introduction

Scalable coding is an important tool for efficient transmission of video over packet switched network. In a scalable coder, essential information for the video source is transmitted in the base layer, and can be decoded independently to obtain a coarse quality of reconstruction. Supplementary information is transmitted in higher enhancement layers, which, when combined with base layer information, improves the video reconstruction at the decoder. Syntax for scalable coding is provided in H.263+ and MPEG standards.

Scalable video coding offers means for robustness as base-layer reconstruction may be used as a fall-back option in case of severe packet loss [1] [2]. For example, ATM networks can assign higher priority in transportation to the base-layer cells in case of congestion. In wireless networks, base-layer packets may be protected by stronger error correction codes than enhancement-layer packets. However, in practice, some packet loss is inevitable even in the base-layer. Moreover, error propagation will amplify the effect of packet losses in both base and enhancement layers, and will further degrade the performance. In this paper, we propose an optimal strategy for coding mode selection per macroblock (MB) in both base and enhancement layers, which substantially improves the robustness of scalable video coding systems. While there is a considerable volume of published work on mode selection for packet loss resilience in the single-layer (non-scalable) video coding (e.g., [3] [4] [5] [6]), very little work has been reported on the corresponding problem in scalable video coding.

We focus on an SNR scalable system, which provides layers with the same spatial-temporal resolution but different reconstruction quality. The key step in our derivation is the estimation of the overall decoder distortion that takes into account the quantization, packet loss, and the error concealment scheme. To calculate this estimate, we extend the recursive

optimal per-pixel estimate (ROPE) which we had proposed for non-scalable video coding [5] [6]. The extended ROPE is shown to accurately account for both temporal and spatial error propagation, and to compute the total distortion in each layer at pixel-level precision. For each MB, the prediction mode and quantization step size are jointly selected to minimize the rate-distortion (RD) cost. Simulation results show substantial gains in reconstructed video PSNR at the base as well as enhancement layers.

The paper is organized as follows. In section II, we derive the extended ROPE model that computes the optimal estimate of the overall distortion of decoder reconstruction for each layer. We incorporate the estimate within an RD framework for optimal selection of mode and quantizer parameter in section III. Section IV presents simulation results to demonstrate the performance of the method.

II. Recursive Optimal per-Pixel Estimate of Decoder Distortion in Scalable Coding

A. Preliminaries

In the standard video coder, the video frame is segmented into MBs. In the base layer, the MBs may be encoded in either inter-mode or intra-mode. In inter-mode, the MB is "predicted" from the previously decoded frame via motion compensation, and the prediction error is encoded. In intra-mode, the original MB data is encoded directly. In the enhancement layer, there are three possible prediction modes [7]. MBs can be predicted from the current base layer (upward), from the previous enhancement layer (forward), or via combined prediction using both (bi-directional). The prediction residue is then transform coded.

Mode selection is a powerful standard compatible tool to trade compression efficiency for packet loss resilience. The use of intra-mode in the base layer, and upward prediction in the enhancement layer, can limit error propagation and is more effective during scene changes. However, in general they require more bits for quantization. An optimal mode selection strategy at the encoder should minimize the overall distortion in decoder reconstruction, which includes the effects of quantization and packet loss, for the given bit rate. Thus, a key task at the encoder is the estimation of overall decoder distortion.

However, this task is complicated by two factors. Spatial error propagation beyond MB boundaries (due to motion compensation) can only be accurately accounted for by computing the distortion per pixel. Further, distortions due to quantization and packet loss are not additive, but are instead combined in a highly complex fashion to produce the overall distortion. In this section, we derive an algorithm to accurately estimate the total distortion in decoder reconstruction at the different layers of a scalable coder.

We assume that the group of blocks (GOB) in each row is carried in a separate packet, and that the packets are independently decodable. Thus, the pixel loss rate equals the packet loss rate. We model the channel as a Bernoulli process with packet loss rate p_b for the base layer, and packet loss rate p_e for the enhancement layer. Note that this model is assumed for presentation simplicity, and more complex models may be considered as well.

Let f_n^i denote the original value of pixel i in frame n, let $\hat{f}_n^i(b)$ and $\hat{f}_n^i(e)$ denote its encoder reconstruction at the base and enhancement layer respectively. The reconstructed values at the decoder, possibly after error concealment, are denoted by $\tilde{f}_n^i(b)$ and $\tilde{f}_n^i(e)$. For the encoder, $\tilde{f}_n^i(b)$ and $\tilde{f}_n^i(e)$ are random variables. Assuming mean square error distortion, the

overall expected distortion for this pixel, at the base and enhancement layers, is given by

$$d_n^i(b) = E\{(f_n^i - \tilde{f}_n^i(b))^2\} = (f_n^i)^2 - 2f_n^i E\{\tilde{f}_n^i(b)\} + E\{(\tilde{f}_n^i(b))^2\}. \quad (1)$$

$$d_n^i(e) = E\{(f_n^i - \tilde{f}_n^i(e))^2\} = (f_n^i)^2 - 2f_n^i E\{\tilde{f}_n^i(e)\} + E\{(\tilde{f}_n^i(e))^2\}. \quad (2)$$

We observe that the computation of $d_n^i(b)$ and $d_n^i(e)$ requires the first and second moments of the corresponding random variables, and develop recursion functions to sequentially compute these two moments.

B. ROPE for the base layer

It is easy to see that the problem of base layer mode selection is identical to that of non-scalable coding. Thus, the ROPE algorithm derived in [5] [6] may be directly applied for calculating the total decoder distortion. We briefly summarize the algorithm in this subsection.

We assume, for presentation simplicity, that the temporal error concealment technique is in use at the decoder. If the MB containing pixel i is lost, temporal replacement is used for error concealment, i.e., the motion vector of this MB is estimated as the median of the motion vectors of the nearest three MBs in the previous GOB (above). Let the estimated motion vector associate pixel i with pixel k in the previous frame. We thus have $\tilde{f}_n^i(b) = \tilde{f}_{n-1}^k(b)$. The probability of this event is $p_b(1-p_b)$. When the previous GOB is also lost, the estimated motion vector is set to zero, and we have $\tilde{f}_n^i(b) = \tilde{f}_{n-1}^i(b)$, with probability ${p_b}^2$. If the MB is correctly received and has been intra-coded, we have $\tilde{f}_n^i(b) = \hat{f}_n^i(b)$ with probability $(1-p_b)$. Thus, for a pixel in an intra-coded MB,

$$\begin{aligned} E\{\tilde{f}_n^i(b)\} &= (1-p_b)(\hat{f}_n^i(b)) \\ &+ p_b(1-p_b)E\{\tilde{f}_{n-1}^k(b)\} + {p_b}^2 E\{\tilde{f}_{n-1}^i(b)\}, \\ E\{(\tilde{f}_n^i(b))^2\} &= (1-p_b)(\hat{f}_n^i(b))^2 \\ &+ p_b(1-p_b)E\{(\tilde{f}_{n-1}^k(b))^2\} + {p_b}^2 E\{(\tilde{f}_{n-1}^i(b))^2\}. \end{aligned} \quad (3)$$

If an inter-coded MB is correctly received, the decoder has access to the quantized residue, $\hat{e}_n^i(b)$, and the motion vector. Let the motion vector be such that pixel i is predicted from pixel j in the previous frame. The encoder's prediction is given by $\hat{g}_n^i(b) = \hat{f}_{n-1}^j(b)$, and its reconstruction is given by $\hat{f}_n^i(b) = \hat{e}_n^i(b) + \hat{g}_n^i(b)$. The decoder must use its prediction, $\tilde{g}_n^i(b) = \tilde{f}_{n-1}^j(b)$. The corresponding reconstruction is given by $\tilde{f}_n^i(b) = \hat{e}_n^i(b) + \tilde{g}_n^i(b)$, with probability $(1-p_b)$. As the decoder's prediction is not identical to encoder's prediction, error propagation occurs even if the residue is received correctly. Thus, for a pixel in an inter-coded MB,

$$\begin{aligned} E\{\tilde{f}_n^i(b)\} &= (1-p_b)(\hat{e}_n^i(b) + E\{\tilde{g}_n^i(b)\}) \\ &+ p_b(1-p_b)E\{\tilde{f}_{n-1}^k(b)\} + {p_b}^2 E\{\tilde{f}_{n-1}^i(b)\}, \\ E\{(\tilde{f}_n^i(b))^2\} &= (1-p_b)E\{(\hat{e}_n^i(b) + \tilde{g}_n^i(b))^2\} \\ &+ p_b(1-p_b)E\{(\tilde{f}_{n-1}^k(b))^2\} + {p_b}^2 E\{(\tilde{f}_{n-1}^i(b))^2\} \end{aligned} \quad (4)$$

$$
\begin{aligned}
&= (1-p_b)((\hat{e}_n^i(b))^2 + 2\hat{e}_n^i(b)E\{\tilde{g}_n^i(b)\} + E\{(\tilde{g}_n^i(b))^2\}) \\
&+ p_b(1-p_b)E\{(\tilde{f}_{n-1}^k(b))^2\} + {p_b}^2 E\{(\tilde{f}_{n-1}^i(b))^2\}.
\end{aligned}
$$

C. ROPE for the enhancement layer

We now extend the ROPE algorithm to estimate the decoder distortion at the enhancement layers. If an MB in the enhancement layer is lost, the decoder uses the corresponding base-layer block for error concealment.

Let us denote the prediction value at the encoder side as $\hat{g}_n^i(e)$, and that of the decoder side as $\tilde{g}_n^i(e)$. Let the transmitted residue is denoted by $\hat{e}_n^i(e)$. Note that $\hat{g}_n^i(e)$ and $\tilde{g}_n^i(e)$ are not identical. Thus, even if the packet containing the current pixel is received correctly (with probability $(1-p_e)$), the reconstruction at the encoder, $\hat{f}_n^i(e) = \hat{e}_n^i(e) + \hat{g}_n^i(e)$, is different from the reconstruction at the decoder, $\tilde{f}_n^i(e) = \hat{e}_n^i(e) + \tilde{g}_n^i(e)$. Note that $\tilde{f}_n^i(e)$ and $\tilde{g}_n^i(e)$ are random variables to the encoder.

Thus, we have the following recursion functions for the expected moments of $\tilde{f}_n^i(e)$:

$$
\begin{aligned}
E\{\tilde{f}_n^i(e)\} &= (1-p_e)(\hat{e}_n^i(e) + E\{\tilde{g}_n^i(e)\}) \\
&+ p_e E\{\tilde{f}_n^i(b)\} \\
E\{(\tilde{f}_n^i(e))^2\} &= (1-p_e)E\{(\hat{e}_n^i(e) + \tilde{g}_n^i(e))^2\} \\
&+ p_e E\{(\tilde{f}_n^i(b))^2\} \\
&= (1-p_e)((\hat{e}_n^i(e))^2 + 2\hat{e}_n^i(e)E\{\tilde{g}_n^i(e)\} + E\{(\tilde{g}_n^i(e))^2\}) \\
&+ p_e E\{(\tilde{f}_n^i(b))^2\}
\end{aligned} \tag{5}
$$

The expected moments of base layer are calculated as described in the previous section.

Let the motion vector of the MB associate pixel i with pixel j in the previous frame. The prediction, at the encoder and decoder, corresponding to the three prediction modes are given by:

- for upward prediction:

$$
\begin{aligned}
\hat{g}_n^i(e) &= \hat{f}_n^i(b), \\
\tilde{g}_n^i(e) &= \tilde{f}_n^i(b)
\end{aligned} \tag{6}
$$

- for forward prediction:

$$
\begin{aligned}
\hat{g}_n^i(e) &= \hat{f}_{n-1}^j(e), \\
\tilde{g}_n^i(e) &= \tilde{f}_{n-1}^j(e).
\end{aligned} \tag{7}
$$

- for bi-directional prediction:

$$
\begin{aligned}
\hat{g}_n^i(e) &= (\hat{f}_{n-1}^j(e) + \hat{f}_n^i(b))/2, \\
\tilde{g}_n^i(e) &= (\tilde{f}_{n-1}^j(e) + \tilde{f}_n^i(b))/2.
\end{aligned} \tag{8}
$$

We reemphasize that these recursions are performed at the encoder in order to calculate the expected total distortion at the decoder precisely per pixel. While for simplicity the recursions have been derived within a two-layer scalable coding setup, they can be extended in a straightforward manner to compute the total decoder distortion at each layer of a multi-layer scalable video coder.

Note that the estimate is precise for integer-pixel motion estimation. In the half-pixel case, the bilinear interpolation makes the exact computation of the second moment highly complex. The estimate is approximated by the simpler recursion of integer-pixel motion compensation. Further, for bi-directional prediction, we assume

$$E\{\tilde{f}_n^i(b)\tilde{f}_{n-1}^j(e)\} = E\{\tilde{f}_n^i(b)\}E\{\tilde{f}_{n-1}^j(e)\}. \tag{9}$$

Although these approximations are sub-optimal, substantial gains are achieved.

D. Simplified ROPE for the special case of guaranteed base layer

An important practical scenario in scalable video coding is when the base-layer packets are transmitted with guaranteed reception or with negligible packet loss rate. In this case, the decoder reconstruction at the base-layer can be well approximated by the encoder reconstruction, i.e., now $\tilde{f}_n^i(b)$ is not a random variable, but $\tilde{f}_n^i(b) = \hat{f}_n^i(b)$. For this special case, we can use a simplified ROPE to calculate the enhancement-layer distortion. The recursions for the enhancement-layer can be rewritten as:

$$\begin{aligned}
E\{\tilde{f}_n^i(e)\} &= (1-p_e)(\hat{e}_n^i(e) + E\{\tilde{g}_n^i(e)\}) \\
&+ p_e\hat{f}_n^i(b) \\
E\{(\tilde{f}_n^i(e))^2\} &= (1-p_e)E\{(\hat{e}_n^i(e) + \tilde{g}_n^i(e))^2\} \\
&+ p_e(\hat{f}_n^i(b))^2 \\
&= (1-p_e)((\hat{e}_n^i(e))^2 + 2\hat{e}_n^i(e)E\{\tilde{g}_n^i(e)\} + E\{(\tilde{g}_n^i(e))^2\}) \\
&+ p_e(\hat{f}_n^i(b))^2
\end{aligned} \tag{10}$$

where the prediction, and , for the three prediction modes are given by:

- for upward prediction:

$$\hat{g}_n^i(e) = \tilde{g}_n^i(e) = \hat{f}_n^i(b). \tag{11}$$

- for forward prediction:

$$\begin{aligned}
\hat{g}_n^i(e) &= \hat{f}_{n-1}^j(e), \\
\tilde{g}_n^i(e) &= \tilde{f}_{n-1}^j(e).
\end{aligned} \tag{12}$$

- for bi-directional prediction:

$$\begin{aligned}
\hat{g}_n^i(e) &= (\hat{f}_{n-1}^j(e) + \hat{f}_n^i(b))/2, \\
\tilde{g}_n^i(e) &= (\tilde{f}_{n-1}^j(e) + \hat{f}_n^i(b))/2.
\end{aligned} \tag{13}$$

III. RD Optimized Mode Selection Algorithm for Scalable Coding

We next incorporate the distortion estimate computed by the ROPE model into an RD framework, and select the coding mode and quantization step size of each MB to minimize the overall decoder distortion for the given bit rate.

The "classical" rate-distortion problem is that of jointly selecting the coding modes for all the MBs to minimize the total distortion, D, subject to a given rate constraint, R. Equivalently, we may recast the problem as an unconstrained Lagrangian minimization, $J = D + \lambda R$, where λ is the Lagrange multiplier. Note that individual MB contributions to this cost are additive and, hence, the cost can be independently minimized for each MB. The coding modes are optimized for the base and enhancement layers sequentially.

For the base layer, the optimal mode and quantization step size for each MB are chosen by the simple minimization:

$$\min_{\text{mode}}(J_{\text{MB}}(b)) = \min_{\text{mode}}(D_{\text{MB}}(b) + \lambda_b R_{\text{MB}}(b)) \tag{14}$$

where the distortion of the MB is the sum of the distortion contributions of the individual pixels:

$$D_{\text{MB}}(b) = \sum_{i \in \text{MB}} d_n^i(b). \tag{15}$$

For the enhancement-layer, the prediction mode and quantization step size are chosen to minimize

$$\min_{\text{mode}}(J_{\text{MB}}(e)) = \min_{\text{mode}}(D_{\text{MB}}(e) + \lambda_e R_{\text{MB}}(e)) \tag{16}$$

where the distortion of the MB is given by:

$$D_{\text{MB}}(e) = \sum_{i \in \text{MB}} d_n^i(e). \tag{17}$$

Note that we use the ROPE model to calculate the distortion per pixel, while the coding mode and quantization step size are selected per MB via (14) and (16). The rate is controlled by using the "buffer status" to update λ_b and λ_e as in [6].

IV. Simulation Results

For the simulations, we implemented the ROPE-RD mode selection strategy by appropriately modifying the UBC H.263+ codec with two-layer scalability [8]. The RTP payload format [9] is assumed for packetization, and each packet contains one GOB. A random packet loss generator is used to drop packets at a specified loss rate. In the proposed system, the ROPE-RD algorithm is used for both layers for selection of mode and quantizer parameter. For comparison, we use random intra-update (RIU) [4] in the base layer, where MBs are randomly intra-coded at the rate of $1/p_b$. In the enhancement layer, we compare the proposed scheme with two standard approaches for prediction mode selection. One method employs the quantization distortion estimate (QDE) within an RD framework to make the selection among the three prediction modes. In the other approach, only the upward prediction (UP) mode is used. UP ensures that there is no error propagation in the base-layer loss free case.

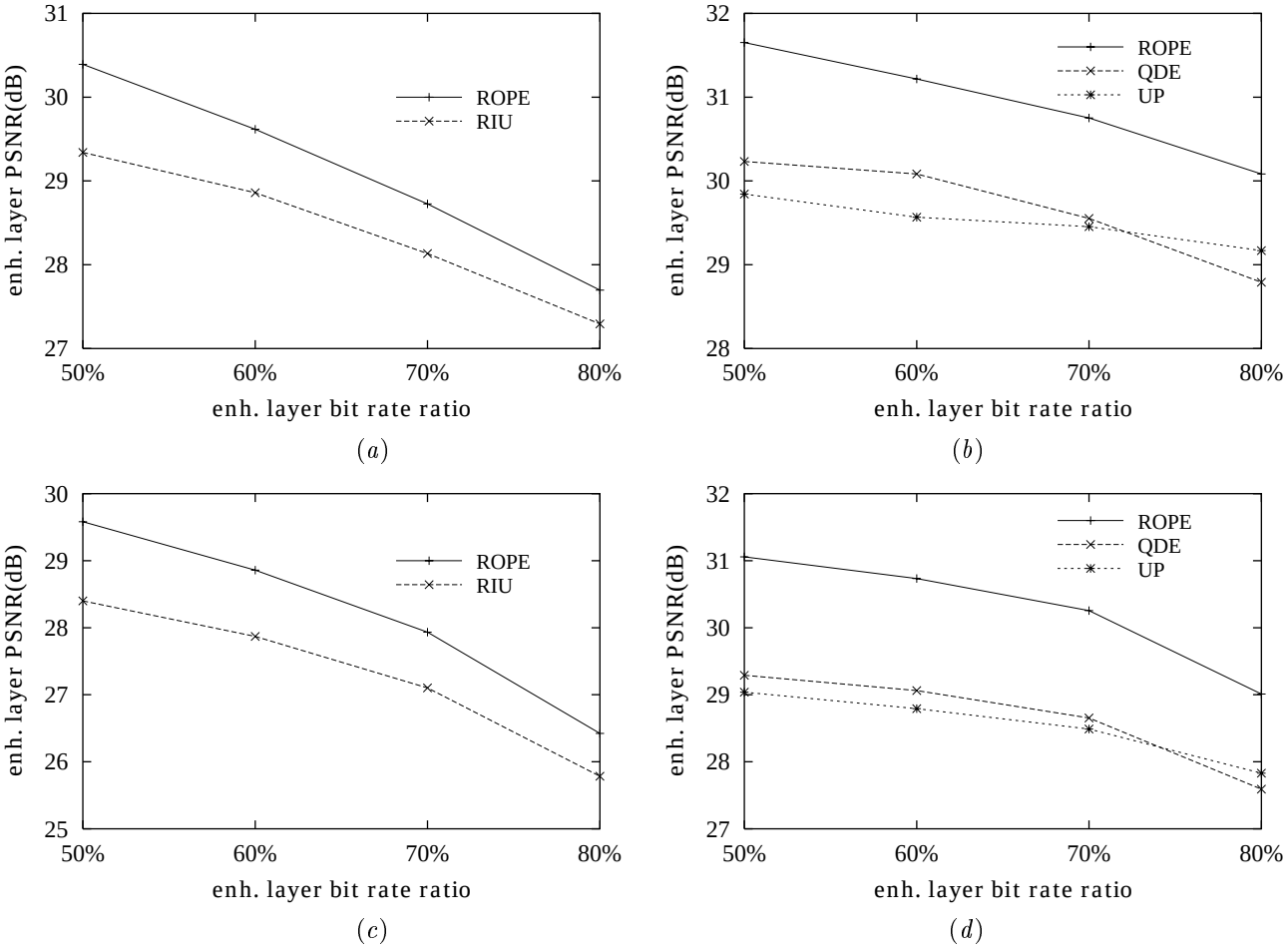

Fig. 1. PSNR vs. enhancement layer bit rate (as a fraction of total rate). Base layer loss prone. Base layer methods: ROPE (proposed), RIU [4]; enhancement layer methods: ROPE (proposed), QDE, UP. Base layer packet loss rate=5%, enhancement layer packet loss rate=15%. QCIF sequence "carphone" (frame rate=10fps, total bit rate=100kbps): (a) base layer PSNR, (b) enhancement layer PSNR. CIF sequence "LTS" (frame rate=15fps, total bit rate=600kbps): (c) base layer PSNR, (d) enhancement layer PSNR.

250 frames from QCIF video sequences "carphone" and CIF video sequence "LTS" are compressed. The PSNR of luminance reconstruction is computed for the sequence and averaged over 30 different channel simulations (with different packet loss patterns).

Figure 1 shows the results for the QCIF sequence "carphone" and CIF sequence "LTS" when the packet loss rates in the base and enhancement layer are 5% and 15% respectively. In the base layer, our proposed ROPE based mode selection outperforms the RIU scheme by about 0.4∼1.0dB for "carphone" and 0.6∼1.2dB for "LTS". In the enhancement-layer, ROPE based robust mode selection achieves PSNR gains of 0.9∼1.8 dB for the "carphone" sequence and 1.2∼2 dB for the "LTS" sequence over the competing methods. This corresponds to additional improvement of 0.5∼0.8dB.

Figure 2 and Figure 3 present the results when reception of base layer packets is guaranteed. In this case, base-layer performance is identical for both the methods of ROPE and RIU. Enhancement layer PSNR is shown versus packet loss rate in Figure 2, and versus

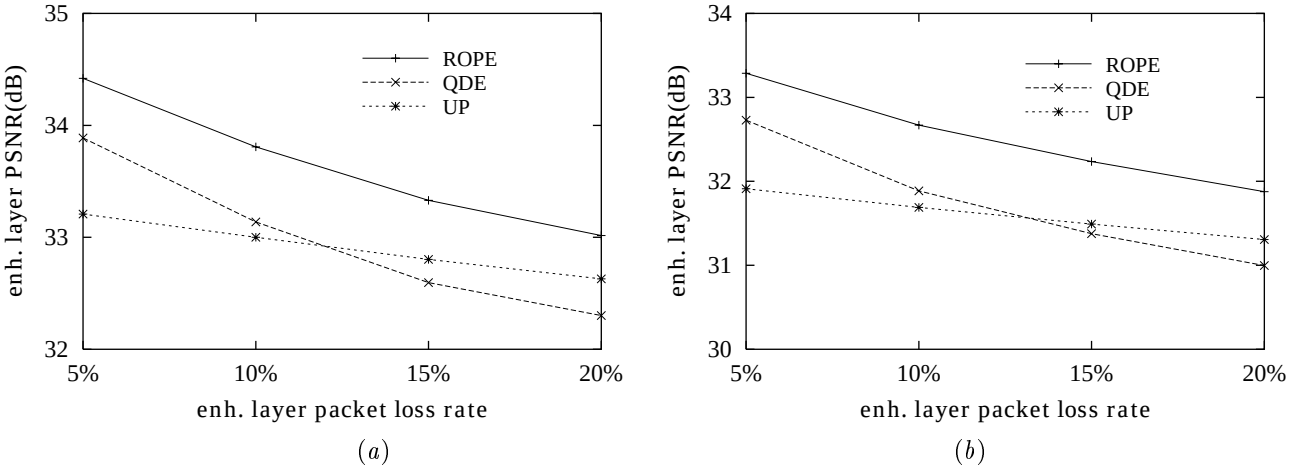

Fig. 2. PSNR vs. enhancement layer packet loss rate. Base layer loss free. Methods: ROPE (proposed), QDE, UP. Enhancement layer bit rate ratio=75%. (a) QCIF sequence "carphone" (frame rate=10fps, total bit rate=100kbps), (b) CIF sequence "LTS" (frame rate=15fps, total bit rate=600kbps).

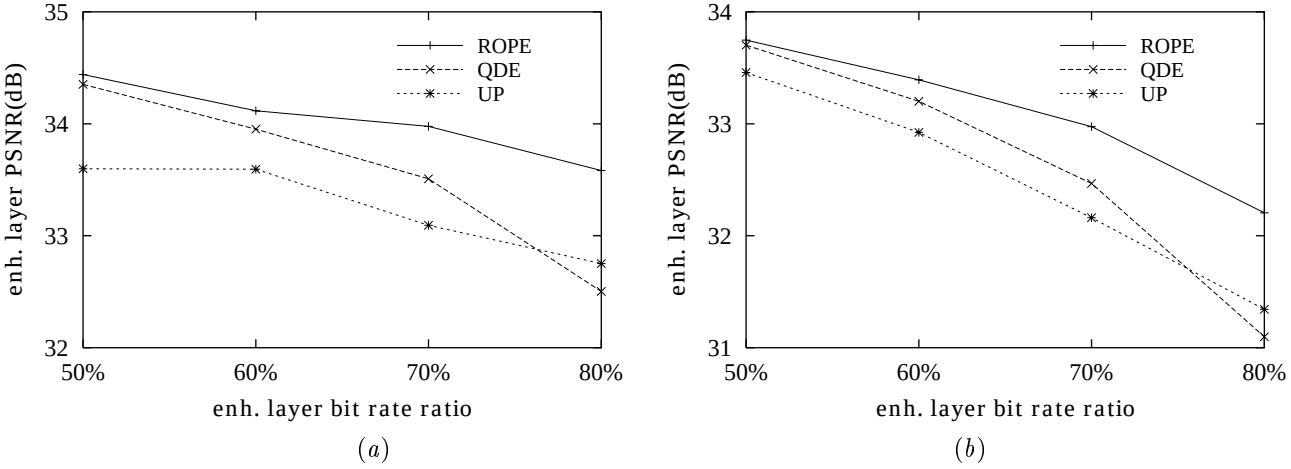

Fig. 3. PSNR vs. enhancement layer bit rate (as a fraction of total bit rate). Base layer loss free. Methods: ROPE (proposed), QDE, UP. Enhancement layer packet loss rate=10%. (a) QCIF sequence "carphone" (frame rate=10fps, total bit rate=100kbps), (b) CIF sequence "LTS" (frame rate=15fps, total bit rate=600kbps).

enhancement layer bit rate (as a fraction of total rate) in Figure 3. Note that the relative performance of QDE and UP depends on the packet loss rate and the enhancement layer bit rate. The proposed ROPE, however, consistently outperforms the other two methods.

Note that similar performance gains can be expected when proposed ROPE-RD mode switching algorithm is incorporated into other scalable video coding schemes such as MPEG.

V. Conclusion

We propose a method for optimal mode selection in scalable video coding, which enhances robustness to packet loss. The method accurately estimates the overall decoder distortion for each layer at pixel-level precision by accounting for quantization, error propagation due

to packet loss, and error concealment scheme employed at the decoder. The estimate is then incorporated within an RD framework for optimal mode selection for macroblocks in each layer. Simulation results show that the proposed method consistently outperforms conventional mode selection methods, and achieves significant PSNR gains in both base and enhancement layer. The algorithm requires no change to the coding syntax or to the decoder. Thus, it is compatible with standards such as H.263+ and MPEG.

Acknowledgement

This work was supported in part by the National Science Foundation under grant MIP-9707764, the University of California MICRO program, Cisco Systems, Inc., Conexant Systems, Inc., Dialogic Corp., Fujitsu Laboratories of America, Inc., General Electric Co., Hughes Network Systems, Lernout & Hauspie Speech Products, Lockheed Martin, Lucent Technologies, Inc., Qualcomm, Inc., and Texas Instruments, Inc.

References

[1] R. Aravind, M. R. Civanlar and A. R. Reibman, "Packet loss resilience of MPEG-2 scalable video coding algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.6, no. 5, Oct. 1996, pp. 426-435.

[2] J. Villasenor, Y.-Q. Zhang, and J.-T. Wen, "Robust video coding algorithms and systems," *Proceedings of the IEEE*, vol.87, no.10, Oct. 1999. pp.1724-33.

[3] E. Steinbach, N. Farber and B. Girod, "Standard compatible extension of H.263 for robust video transmission in mobile environments," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol.7, No.6, Dec. 1997, pp. 872-881.

[4] G. Cote and F. Kossentini, "Optimal intra coding of blocks for robust video communication over the Internet," *Signal Processing: Image Communication*, vol.15, No. 1-2, Sept. 1999, pp. 25-34.

[5] R. Zhang, S. L. Regunathan and K. Rose, "Optimal intra/inter mode switching for robust video communication over the Internet," *Thirty-third Asilomar Conference on Signals, Systems, and Computers*, Oct. 24-29, 1999.

[6] R. Zhang, S. L. Regunathan and K. Rose, "Video coding with optimal intra/inter mode switching for packet loss resilience," to appear on *IEEE Journal of Selected Areas in Communications, special issue on Error-Resilient Image and Video Transmission.*

[7] ITU-T Recommendation H.263, "Video coding for low bit rate communication," 1998

[8] H.263+ codec, http://spmg.ece.ubc.ca/

[9] "RTP Payload Format for the 1998 Version of ITU-T Rec. H.263 Video (H.263+)," Internet Draft, *RFC2429,ftp://ftp.isi.edu/in-notes/rfc2429.txt*