Optimal Intra/Inter Mode Switching for Robust Video Communication over the Internet

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

Abstract

Robustness to packet loss is a critical requirement in video coding for transmission over the Internet. We propose an intra/inter mode switching scheme to stop error propagation, while optimizing compression efficiency. The encoder computes the expected distortion at the decoder precisely ***per pixel****, while accounting for quantization, packet loss rate, and effect of error concealment. The distortion estimate is incorporated within an overall rate-distortion framework to optimally select the coding mode of each macroblock. Simulation results demonstrate the accuracy of the estimate, and show that substantial gains in PSNR can be achieved over state-of-the-art RD and non-RD based mode-switching approaches for transmission over networks with packet loss.*

1. Introduction

In packet-switched networks, packets may be discarded due to buffer overflow at intermediate nodes, or be considered lost due to long queuing delays. For example, the packet loss rate in Internet communications may reach the level of 20% [1]. The problem is exacerbated in the case of (standard) predictive video coding where the prediction loop propagates errors and causes substantial, and sometimes catastrophic, deterioration of the received video signal.

Intra-coding is an important tool for mitigating the effects of packet loss. By switching off the prediction for certain macro-blocks, error propagation is stopped. Intra-coding requires no modification to the bitstream syntax, and is, hence, standard-compatible. However, intra-coding typically costs more in rate than inter-coding. An important open problem is that of mode switching for each MB, so as to optimize the tradeoff between the compression efficiency and robustness. Periodic intra-coding of whole frames [2], contiguous blocks [1], or random blocks [3] using a heuristic refresh frequency have been proposed. Other methods apply frequent intra-update to regions that undergo significant changes [4], or where a rough estimate of the decoder error exceeds a given threshold [5][6]. A more direct solution incorporates mode selection within an overall rate-distortion framework. An early proposal of rate-distortion (RD) based mode selection to combat packet loss appeared in [7]. A significant improvement to RD based mode selection was proposed in [3][8] where the the encoder takes into account the effects of error concealment. Refer to [11] for a more general discussion of RD frameworks that incorporate channel error.

Although RD based mode selection methods [3][7][8] represent a significant advance over heuristic mode switching strategies, they suffer from a severe drawback. The encoders in these schemes do not possess the capability to *accurately* estimate the overall distortion (due to quantization, concealment and error propagation) in the decoder frame reconstruction. In [7] simple approximative distortion estimation is suggested, while the encoder in [3][8] ignores error propagation beyond one frame and approximates the total block distortion as a sum of the quantization distortion of that block, and weighted concealment distortion of corresponding blocks in the previous frame.

The main contribution of our work is a method to *optimally estimate*, at the encoder, the overall distortion of the decoder. The method uses a *recursive* algorithm to estimate the total distortion at *pixel level precision*, and thus accurately account for error propagation along both the temporal and spatial axes. We demonstrate the accuracy of the estimate through simulation results and, further, provide compelling experimental evidence that incorporation of the optimal estimator within an RD based mode switching algorithm achieves substantial performance gains over state-of-the-art RD and non-RD based mode switching algorithms.

In section 2, we derive an algorithm that computes the optimal estimate of the overall distortion of decoder reconstruction precisely per pixel. We incorporate the estimator

0-7803-5700-0/99/$10.00©1999 IEEE

within an RD framework for optimal mode switching in section 3. Section 4 presents simulation results to demonstrate the performance of the method.

2. Recursive Optimal Per-Pixel Estimate of Decoder Distortion

The standard video coder employs inter-frame prediction to remove temporal redundancies, and transform coding to exploit spatial redundancies. The video frame is segmented into "macroblocks" (MBs) that are encoded either in 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 coded. In intra-mode, the original MB data is coded directly. Although inter-mode generally achieves better compression, it promotes error propagation. Note that motion compensation leads to spatial error propagation beyond MB boundaries. Hence, only by computing the distortion per pixel can we accurately account for error propagation, and truly optimize the mode switching strategy. Further, note that the distortion due to quantization and concealment are not simply additive. Instead, they are combined in a highly complex fashion to produce the overall distortion. We now derive a method to optimally estimate the total decoder distortion for the given rate, packet loss condition and error concealment method.

We assume that the group of blocks(GOB) in each row are carried in a separate packet. The packets are independently decodable and the pixel loss rate equals the packet loss rate. We model the channel as a Bernoulli process with packet loss rate p. If a packet is lost, the decoder performs temporal replacement for error concealment as follows: The motion vector of a lost MB is estimated as the median of the motion vectors of the nearest three MBs in the previous GOB (above). When the previous GOB is also lost, the estimated motion vector is set to zero. The missing pixels are replaced by the corresponding pixels in the previous frame.

Let f_n^i denote the original value of pixel i in frame n and $\hat{f}_n^i$ denote its encoder reconstruction. The reconstructed value at the decoder, possibly after error concealment, is denoted by $\tilde{f}_n^i$. For the encoder, $\tilde{f}_n^i$ is a random variable. Using the mean square error as distortion metric, the overall expected distortion for this pixel is

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

We observe that the computation of d_n^i requires the first and second moments of each random variable in the sequence $\tilde{f}_n^i$. We develop recursion functions to sequentially compute these two moments. We consider two cases depending on whether the pixel belongs to an intra-coded MB or an inter-coded MB.

Pixel in an intra-coded (*I*) MB: If the packet containing this MB received correctly, we have $\tilde{f}_n^i = \hat{f}_n^i$. The probability of this event is $1 - p$. If packet is lost and the previous GOB is available, the median motion vector of the nearest three MBs is calculated and used to associate pixel i in the current frame with pixel k in the previous frame. We thus have $\tilde{f}_n^i = \tilde{f}_{n-1}^k$ with probability $p(1-p)$. If the previous GOB is also lost, the motion vector estimate is set to zero, and we have $\tilde{f}_n^i = \tilde{f}_{n-1}^i$ with probability p^2. Thus,

$$\begin{aligned} E\{\tilde{f}_n^i\}(I) &= (1-p)(\hat{f}_n^i) \\ &+ p(1-p)E\{\tilde{f}_{n-1}^k\} + p^2 E\{\tilde{f}_{n-1}^i\}, \qquad (2) \\ E\{(\tilde{f}_n^i)^2\}(I) &= (1-p)(\hat{f}_n^i)^2 \\ &+ p(1-p)E\{(\tilde{f}_{n-1}^k)^2\} + p^2 E\{(\tilde{f}_{n-1}^i)^2\}. \end{aligned}$$

Pixel in an inter-coded (*I*) MB: Let the motion vector of the MB be such that pixel i is predicted from pixel j in the previous frame. Thus the encoder prediction is $\hat{f}_{n-1}^j$. Let the quantized prediction error be denoted by $\hat{e}_n^i$. Since the encoder reconstruction of this pixel, $\hat{f}_n^i$, is obtained by adding the quantized residue to the prediction, we have $\hat{e}_n^i = \hat{f}_n^i - \hat{f}_{n-1}^j$. If the packet is currently received, the decoder has access to both $\hat{e}_n^i$ and the motion vector. But, it must use for prediction the *decoder's* reconstruction of pixel j in the previous frame, $\tilde{f}_{n-1}^j$. Thus the decoder reconstruction of pixel i is given by $\tilde{f}_n^i = \hat{e}_n^i + \tilde{f}_{n-1}^j$ with probability $1 - p$. This explains how the error propagates even if the packet is received currently. If the packet is lost, the error concealment is performed in a manner identical to intra-coded MB. Thus,

$$\begin{aligned} E\{\tilde{f}_n^i\}(P) &= (1-p)(\hat{e}_n^i + E\{\tilde{f}_{n-1}^j\}) \\ &+ p(1-p)E\{\tilde{f}_{n-1}^k\} + p^2 E\{\tilde{f}_{n-1}^i\}, \\ E\{(\tilde{f}_n^i)^2\}(P) &= (1-p)E\{(\hat{e}_n^i + \tilde{f}_{n-1}^j)^2\} \\ &+ p(1-p)E\{(\tilde{f}_{n-1}^k)^2\} \\ &+ p^2 E\{(\tilde{f}_{n-1}^i)^2\}. \qquad (3) \end{aligned}$$

We reemphasize that these recursions are performed at the *encoder* in order to calculate the expected distortion at the *decoder* precisely per pixel. The estimate is precise for integer-pixel motion estimation. For the half-pixel case, the bilinear interpolation makes the exact computation of the second moment highly complex. However, we found the estimate is well approximated by the simpler recursion of integer-pixel motion compensation and, although strictly speaking it is sub-optimal, substantial gains are maintained.

We now discuss the accuracy of the proposed "recursive optimal per-pixel estimate"(ROPE) . We compare our estimate with the approach recently proposed in [3][8], which we will refer to as the "block-weighted distortion estimate"(BWDE). BWDE estimates the decoder distortion via the formula $\hat{D} = pD_c + (1-p)D_q$, where D_c of a block is defined as the weighted average of concealment distortion of the previous frame blocks that are mapped to it by motion compensation, and D_q is the quantization distortion of the current block [12]. Note that this approach assumes that the distortion is additive in its concealment and quantization components. As an additional reference, we use the simplistic estimate based only on the quantization distortion, which we call the "quantization distortion estimate"(QDE).

We implemented the three approaches by modifying the H.263 codec [14]. In the simulation, we select MBs for intra-updates at period intervals and estimate the decoder frame distortion by the above three methods. Figure 1 compares these estimates with the actual decoder distortion averaged over 30 channel realizations (with different packet loss patterns). It is evident that the proposed ROPE model provides a highly accurate estimate of the decoder distortion, and substantially outperforms its competitors. This advantage is obtained at the cost of a modest increase in computational complexity of the encoder.

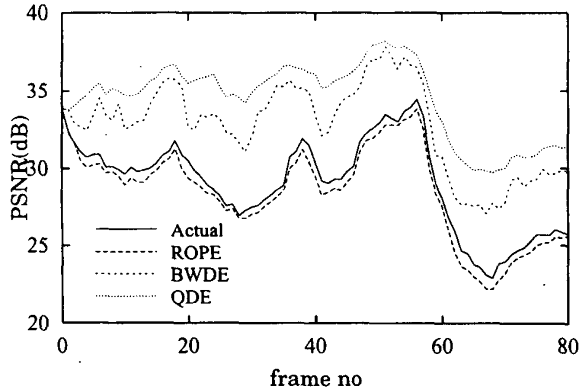

Figure 1. Comparison between actual and estimated decoder PSNR in the integer pixel motion compensation case. Competing estimators: ROPE (proposed), BWDE [3][8], QDE. "carphone", r=100kbps, f=10fps, p=10%

3. RD based Mode Switching Algorithm

The ROPE-RD system incorporates the distortion estimate computed by the ROPE model into an RD framework, and thus selects the coding mode of each MB to minimize the overall distortion for the given bit rate.

The "classical" rate-distortion problem is that of switching between the coding modes per MB to minimize the total distortion, D, subject to a given rate constraint, R. Equivalently, we may recast the problem as an unconstrained Lagrange minimization, $J = D + \lambda R$, where λ is the Lagrangian multiplier. Note that individual MB contributions to this cost are additive and, hence, the cost can be independently minimized for each MB. Therefore, the optimal encoding mode for each MB is chosen by a simple minimization:

$$\min_{\text{mode}}(J_{\text{MB}}) = \min_{\text{mode}}(D_{\text{MB}} + \lambda R_{\text{MB}}), \qquad (4)$$

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

$$D_{\text{MB}} = \sum_{i \in \text{MB}} d_n^i. \qquad (5)$$

Note that we use the ROPE model to calculate the distortion *per pixel*, while the coding mode is selected *per MB* via (4). The rate is controlled by using the "buffer status" to update λ via

$$\lambda_{n+1} = \lambda_n(1 + \alpha(\sum_{i=1}^{n} R_i - nR_{\text{target}})), \qquad (6)$$

where α is given by

$$\alpha = \frac{1}{5R_{\text{target}}}. \qquad (7)$$

For each MB, the *mode* and the *quantization step size* are selected to minimize the rate-distortion Lagrangian.

Extension to Motion Vector Optimization: As is the case for error free channels [13], the performance of the ROPE-RD system can be further improved by incorporating the selection of motion vectors within the RD framework. A small number of motion vector candidates are preselected from the conventional motion estimation step via block matching. From these candidates, the motion vector for the MB is selected jointly along with the coding mode and the quantization step size by optimizing equation 4.

4. Simulation Results

We implemented the ROPE-RD mode switching strategy by appropriately modifying the Telenor H.263 codec [14]. The RTP payload format [15] is used for packetization. A random packet loss generator is used to drop packets at the specified loss rate. We compare the proposed ROPE-RD mode switching algorithm with three known schemes: BWDE-RD [3][8], "Scattered-Block Intra Update" (SB-IU) and "Contiguous-Block Intra Update(CB-IU)" [1]. BWDE-RD [3][8] incorporates BWDE into an RD framework. The

SB-IU method assigns the MBs to $1/p$ groups, and updates one group per frame so that the intra updating frequency for each MB is $1/p$ [3]. The CB-IU method follows the suggestion in [1], where contiguous-block patterns are recommended depending on the packet loss rate. Temporal-replacement is used for error concealment in all the coding methods. The average PSNR of the decoder reconstruction is computed over 30 channel realizations (with different packet loss patterns).

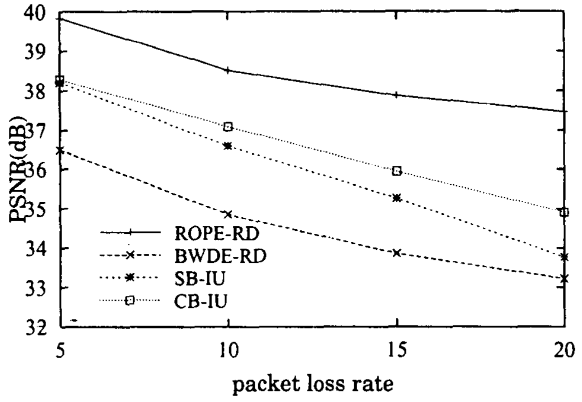

Figure 2. PSNR vs. packet loss rate. Methods: ROPE-RD (proposed), BWDE-RD [3][8], SB-IU [3], CB-IU [1]."salesman", half pixel motion compensation, r=300kbps, f=30fps.

Table 1 and Table 2 summarize the simulation results of encoding 6 QCIF sequences. We use 150 frames in *miss_america* and 250 frames in the other sequences. Table 1 shows the results for the case of bit rate of 100kbps and frame rate of 10fps, while table 2 presents the corresponding results for bit rate of 300kbps and frame rate of 30fps. The packet loss rate is 10% in both cases. Figure 2 shows the performance versus packet loss rate for all the methods on the *salesman* sequence. The results demonstrate that precise distortion estimation enables the ROPE-RD algorithm to achieve consistent and significant gains over known RD and non-RD based mode switching algorithms.

Figure 3 illustrates the advantage of incorporating the selection of motion vectors within the RD framework. PSNR versus packet loss rate at different bit rates is shown for ROPE-RD method without motion vector optimization (1 motion vector candidate), and ROPE-RD method with motion vector optimization (3 motion vector candidates). Motion vector optimization is seen to yield small additional improvements (0.2-0.5dB) in performance at very low bit rates.

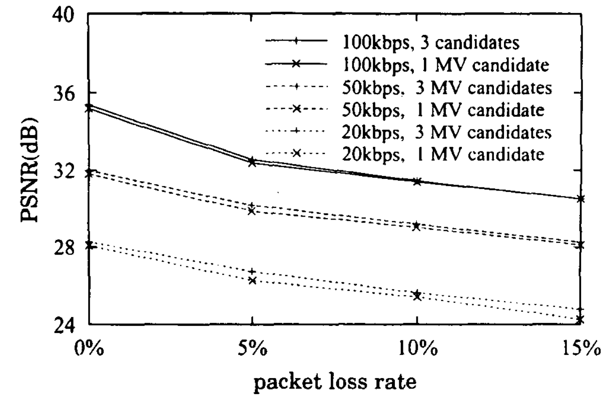

Figure 3. PSNR vs. packet loss rate at different bit rate. Methods: ROPE-RD without motion vector optimization (1 MV candidate), ROPE-RD with motion vector optimization (3 MV candidates). "carphone", integer pixel motion compensation, f=10fps.

5. Conclusion

We proposed a method for rate distortion optimized intra/inter mode switching, which enhances the robustness of video coders to packet loss. The method accurately estimates the total decoder distortion at pixel-level precision by accounting for quantization, concealment and error propagation. Simulation results show that the proposed method substantially and consistently outperforms state-of-the-art RD and non-RD based mode switching methods. Motion vector optimization within RD framework yields small additional gains. The proposed method is standard-compatible.

References

[1] Q. F. Zhu and L. Kerofsky, "Joint source coding, transport processing and error concealment for H.323-based packet video," *VCIP 99*, San Jose CA, pp. 52-62, Jan. 1999.

[2] T. Turletti and C. Huitema, "Videoconferencing on the Internet," *IEEE/ACM Transactions on Networking*, pp. 340-351, Vol.4, No.3, Jun.1996.

[3] G. Cote and F. Kossentini,"Optimal Intra Coding of Blocks for Robust Video Communication over the Internet," in *the special issue on "Real-time Video over the Internet", Image Communication*, Aug.1999

Table 1. Performance comparison on QCIF sequences. integer pixel motion compensation, r=100kbps, f=10fps, p=10%.

Sequence	ROPE-RD	BWDE-RD	SB-IU	CB-IU
Miss America	38.96dB	38.05dB	38.38dB	36.23dB
Grandma	36.94dB	35.48dB	35.20dB	34.93dB
Salesman	35.36dB	33.35dB	32.28dB	32.30dB
Mother/Daughter	34.10dB	31.62dB	32.51dB	31.96dB
Carphone	31.49dB	30.18dB	28.89dB	28.25dB
Foreman	27.98dB	26.90dB	25.02dB	24.61dB

Table 2. Performance comparison on QCIF sequences. integer pixel motion compensation, r=300kbps, f=30fps, p=10%.

Sequence	ROPE-RD	BWDE-RD	SB-IU	CB-IU
Miss America	42.44dB	39.65dB	39.17dB	39.21dB
Grandma	40.20dB	36.23dB	38.64dB	38.99dB
Salesman	38.54dB	34.43dB	36.16dB	36.63dB
Mother/Daughter	37.59dB	31.66dB	36.86dB	36.98dB
Carphone	33.81dB	30.68dB	32.99dB	32.71dB
Foreman	31.58dB	28.88dB	30.83dB	30.52dB

[4] P. Haskell and D. Messerschmitt, "Resynchronization of motion compensated video affected by ATM cell loss," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing, ICASSP'92*, San Francisco, vol.3, pp. 545-548, Mar. 1992.

[5] J. Y. Liao and J. D. Villasenor, "Adaptive intra update for video coding over noisy channels," *ICIP96*, pp. 763-766.

[6] 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*, pp. 872-881, Vol.7, No.6, Dec. 1997.

[7] R. O. Hinds, T. N. Pappas and J. S. Lim, "Joint block-based video source/channel coding for packet-switched networks," *in SPIE*, vol. 3309, pp. 124-133, 1998.

[8] S. Wenger and G. Cote, "Using RFC2429 and H.263+ at low to medium bit-rates for low-latency applications," *Packet Video Workshop 99*, New York City, Apr.1999, http://spmg.ece.ubc.ca/pub/pvw99

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

[10] C. Hsu, A. Ortega and M. Khansari, "Rate control for robust video transmission over wireless channels," *Proc. Visual Communications and Image Processing VCIP'97*, San Jose, SPIE vol. 3024, pp/ 1200-1211, Feb. 1997.

[11] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23-50, Nov. 1998.

[12] F. Kossentini, Private communication, Sep. 1999.

[13] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression,"*IEEE Signal Processing Magazine* ,vol. 15, no. 6, pp. 74-90, Nov. 1998.

[14] ftp://bonde.nta.no/pub/tmn/software.

[15] "RTP Payload Format for the 1998 Version of ITU-T Rec. H.263 Video (H.263+)" Internet Draft, RFC2429, http://www.faqs.org/rfcs/rfc2429.html