Video Coding with Optimal Inter/Intra-Mode Switching for Packet Loss Resilience

Rui Zhang, *Student Member, IEEE*, Shankar L. Regunathan, *Student Member, IEEE*, and Kenneth Rose, *Member, IEEE*

***Abstract*—Resilience to packet loss is a critical requirement in predictive video coding for transmission over packet-switched networks, since the prediction loop propagates errors and causes substantial degradation in video quality. This work proposes an algorithm to optimally estimate the *overall* distortion of decoder frame reconstruction due to quantization, error propagation, and error concealment. The method *recursively* computes the total decoder distortion at *pixel level precision* to accurately account for spatial and temporal error propagation. The accuracy of the estimate is demonstrated via simulation results. The estimate is integrated into a rate-distortion (RD)-based framework for optimal switching between intra-coding and inter-coding modes per macroblock. The cost in computational complexity is modest. The framework is further extended to optimally exploit feedback/acknowledgment information from the receiver/network. Simulation results both with and without a feedback channel demonstrate that precise distortion estimation enables the coder to achieve substantial and consistent gains in PSNR over known state-of-the-art RD- and non-RD-based mode switching methods.**

***Index Terms*—Intra-mode, intra-refresh, packet loss, packet video, rate-distortion, video coding.**

I. Introduction

VIDEO signals have traditionally been transmitted over networks that provide a guaranteed Quality of Service (QoS) for the connection. Therefore, the focus of video coding has been almost exclusively concerned with compression efficiency. In recent years, however, packet-switched networks such as the Internet have become overwhelmingly important. Some of these networks currently provide limited or no end-to-end QoS guarantees. Further, it is anticipated that wireless extensions to the wired backbone will result in additional QoS bottlenecks. Consequently, research on video coder design for packet-switched networks is facing major new challenges.

In packet-switched networks, packets may be discarded due to buffer overflow at intermediate nodes of the network, or may be considered lost due to long queuing delays. This problem is severe, and the packet loss rate in Internet communications, for example, may reach 20% [1]. Clearly, robustness to packet loss is a crucial requirement. 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.

A variety of techniques have been proposed to enhance the robustness of the video communication system to packet loss [2]–[11]. It is widely recognized that intra-coding is an important tool for mitigating the effects of packet loss. By switching off the inter-frame prediction loop for certain macroblocks (MB's), the reproduced blocks are no longer dependent on past frames and error propagation is stopped. Further, intra-coding requires no modification to the bit stream syntax and is, hence, compatible with standards such as H.263+. However, the robustness provided by intra-coding may be costly, as it typically requires a higher bit rate than inter-coding (with prediction). Too many intra-coded MB's will significantly degrade the compression performance. Thus, the problem of switching between intra-coding and inter-coding, so as to achieve the right balance between compression efficiency and robustness, is very important and has been widely addressed. Periodic intra-coding of whole frames [12], or contiguous blocks [1], or random blocks [13] has been proposed. These methods use a heuristic relationship between the packet loss rate and the refresh frequency, but apply intra-coding uniformly to all the regions of the frame. "Content adaptive" methods apply frequent intra-update to regions that undergo significant changes [14], or where a rough estimate of decoder error exceeds a given threshold [15], [16]. A more direct and complete solution incorporates mode selection within an overall rate-distortion (RD) framework so as to directly optimize the performance. An early proposal of mode selection based on an RD framework to combat packet loss appeared in [17]. A significant improvement to RD based mode selection was proposed in [13] and [18] where the encoder takes into account the effects of error concealment. Here, the encoder can give priority to intra-coding of regions where the effect of packet loss is more severe. Incorporation of error concealment in computation of decoder distortion at the encoder has also been described in [10, Appendix 1, Error Tracking], [16], and [19]. Refer also to [20] for a more general discussion of RD frameworks that incorporate channel error.

Although RD-based mode selection methods [17], [13], [18] represent a significant advance over the early 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 in the decoder frame reconstruction (which is due to quantization and concealment after packet loss and error propagation). In [17] simple approximative distortion estimation is suggested, while the encoder in [13]

Manuscript received May 10, 1999; revised October 25, 1999. This work was supported in part by 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.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: rose@ece.ucsb.edu).

Publisher Item Identifier S 0733-8716(00)04342-0.

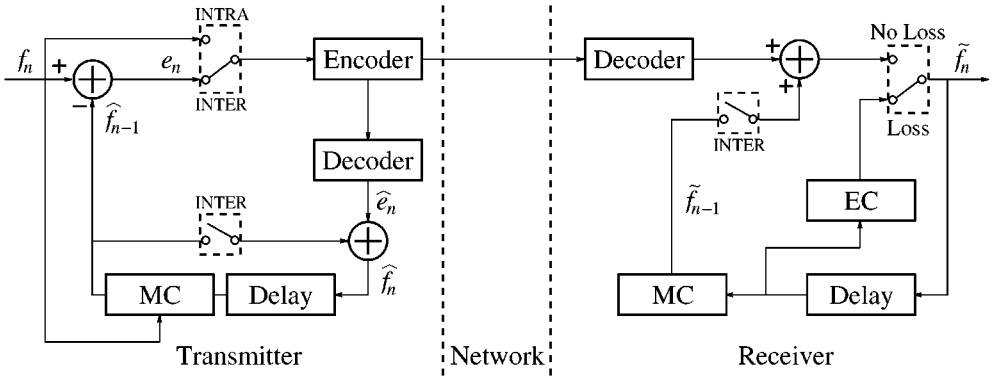


Fig. 1. Predictive video coding with inter/intra-mode switches. (MC: motion compensation, EC: error concealment.)

and [18] ignores error propagation beyond one frame and, further, approximates the total block distortion as a simple 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 for the given rate, packet loss, and error concealment method. 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 (due to motion compensation) 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 [13], [18] and non-RD-based mode [1], [16] switching algorithms.

This paper is organized as follows. In Section II, we analyze the error propagation problem in video coding, and describe the necessary conditions for an accurate decoder distortion estimate. In Section III, we derive an algorithm that computes, at the encoder, the optimal estimate of the overall distortion of decoder reconstruction at *pixel level precision*. The accuracy of the estimate is demonstrated via simulation results. We incorporate the estimate within a rate-distortion based optimal mode switching coder in Section IV, and present results to demonstrate the performance of the method. In Section V, we extend our framework to the scenario where a feedback channel is available. The superiority of the proposed approach over other state-of-the-art mode switching techniques that use feedback is demonstrated by simulation.

II. Error Propagation and Overall Distortion

The common video coding scheme is hybrid and employs inter-frame prediction to remove temporal redundancies, and (typically, discrete cosine) transform coding to exploit spatial redundancies. The video frame is segmented into "macroblocks" (MB's) that are sequentially encoded. Each MB may be encoded in one of two coding modes: inter-mode and intra-mode. In inter-mode, the MB is first "predicted" from the previously decoded frame via motion compensation. Then the prediction error, or residue, is transform-coded. In intra-mode, the original MB data are transform-coded directly without recourse to prediction. Although operation in inter-mode generally achieves higher compression efficiency, it is more sensitive to channel errors as it promotes error propagation. To further illustrate this point, let us refer to the video communication system of Fig. 1.

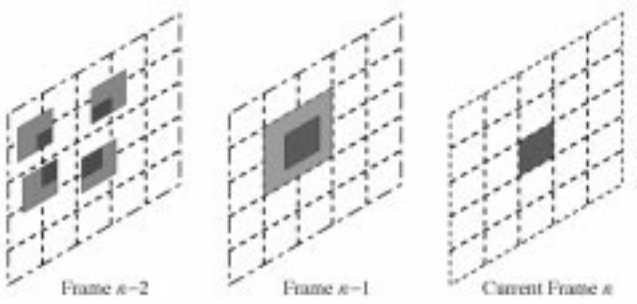


Fig. 2. Origins of pixels in a current block, and the effect of motion compensation on spatial and temporal error propagation.

Let a packet containing data from the current frame be lost in the channel, and let the decoder perform error concealment. Clearly, the resulting reconstruction at the decoder is different from the reconstruction at the encoder. Now, if inter-mode coding is employed to encode the next frame, errors will propagate to it via the prediction step. Whenever the motion vector is nonzero, the error propagates in both the temporal and the spatial directions. Such propagation will continue until an intra-coded block, which is independent of prior frames, is correctly received.

As shown in Fig. 2, the pixels in the current MB may have been motion compensated from pixels in different MB's in the previous frame, each with potentially different error propagation history. Clearly, motion compensation leads to spatial error propagation beyond MB boundaries. Hence, only by computing the estimate of the decoder reconstruction of each individual pixel can we accurately account for error propagation, and truly optimize the mode switching strategy. Further, note that for virtually all useful distortion measures, including the mean square

error criterion, the distortion due to quantization and the distortion due to concealment are not additive. Instead, they are combined in a highly complex fashion to produce the overall distortion. In the next section, we derive a method to optimally estimate the total decoder distortion for the given rate, packet loss condition and error concealment method. The distortion is calculated for *each pixel* in a frame via a simple recursion to accurately account for both spatial and temporal error propagation. The optimal estimate is incorporated within an RD framework for optimal switching between inter-coding and intra-coding per MB. We show that performance of RD-based mode switching is substantially improved by the use of the precise distortion estimate.

III. Recursive Optimal per-Pixel Estimate of Decoder Distortion

A. Preliminaries

The packet video system we consider is shown in Fig. 1. Frame n of the original video signal, denoted by f_n, is compressed and the *encoder* reconstruction is $\hat{f}_n$. The bits are packetized and transmitted over the network. The packets are constructed such that each packet can be independently decoded, and, hence, the loss of one packet does not prevent the decoding of other packets (of course, the reconstruction may still suffer from inter-frame error propagation). In our coding system, we form a group of blocks (GOB) from all the MB's in a particular row (slice), and assume that each GOB is carried in a separate packet. In this setting, the loss rate of a pixel equals the packet loss rate p.

We assume that the packet loss rate p is available at the encoder. This can be either specified as part of the initial negotiations, or adaptively calculated from information provided by the transmission protocol. For example, the real time control protocol (RTCP) [2] provides the encoder with information for calculation of packet loss rate, packet delay, and delay jitter.

The packets are decoded at the receiver. When a packet is lost, an error concealment technique is used for estimating the missing video segment. We denote the decoded (and possibly error-concealed) reconstruction of frame n at the receiver by $\tilde{f}_n$. Note that the encoder does not have access to the value of $\tilde{f}_n$ and must treat it as a random variable.

Any of the error concealment techniques that have been proposed in the literature (see [8]) may be used at the decoder. The temporal-replacement method is relatively simple and efficient [13], [14]. We used the temporal-replacement method in our experiments as follows. The motion vector of a missing MB is estimated as the median of motion vectors of the nearest three MB's in the preceding GOB (above). If the previous GOB is lost, too, the estimated motion vector is set to zero. The pixels in the previous frame, that are pointed to by the estimated motion vector, are used to replace the missing pixels in the current frame.

B. Expected Decoder Distortion per Pixel

Let f_n^i denote the original value of pixel i in frame n, and let $\hat{f}_n^i$ denote its *encoder* reconstruction. The reconstructed value at the *decoder*, possibly after error concealment, is denoted by $\tilde{f}_n^i$. Recall that 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

$$\begin{aligned} d_n^i &= E\left\{\left(f_n^i - \tilde{f}_n^i\right)^2\right\} \\ &= (f_n^i)^2 - 2f_n^i E\left\{\tilde{f}_n^i\right\} + E\left\{\left(\tilde{f}_n^i\right)^2\right\}. \end{aligned} \tag{1}$$

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. For the recursion step, we consider two cases depending on whether the pixel belongs to an intra-coded MB or an inter-coded MB.

1) Pixel in an Intra-Coded MB: Let us first assume that the packet containing the intra-coded MB to which the pixel i belongs is received correctly. We thus have $\tilde{f}_n^i = \hat{f}_n^i$, and the probability of this event is $1-p$. If the packet is lost, the decoder first checks if the previous GOB (above) has been received correctly. If the previous GOB is available, the median motion vector of the nearest three MB's is calculated and used to associate pixel i in the current frame with pixel k in the previous frame. We thus have $\tilde{f}_n^i = \tilde{f}_{n-1}^k$, and the probability of this event is $p(1-p)$. On the other hand, if the previous GOB was lost as well, we set the motion vector estimate to zero. Thus, we have $\tilde{f}_n^i = \tilde{f}_{n-1}^i$, with probability p^2. Combining the three cases, the first and second moments of $\tilde{f}_n^i$ for a pixel in an intra-coded MB are given by (intra-mode denoted by I)

$$\begin{aligned} E\left\{\tilde{f}_n^i\right\}(I) &= (1-p)\left(\hat{f}_n^i\right) + p(1-p)E\left\{\tilde{f}_{n-1}^k\right\} \\ &\quad + p^2 E\left\{\tilde{f}_{n-1}^i\right\} \end{aligned} \tag{2}$$

$$\begin{aligned} E\left\{\left(\tilde{f}_n^i\right)^2\right\}(I) &= (1-p)\left(\hat{f}_n^i\right)^2 + p(1-p)E\left\{\left(\tilde{f}_{n-1}^k\right)^2\right\} \\ &\quad + p^2 E\left\{\left(\tilde{f}_{n-1}^i\right)^2\right\}. \end{aligned} \tag{3}$$

2) Pixel in an Inter-Coded MB: The derivation of the moments is more complex if the pixel belongs to an inter-coded MB. Let us assume that the true motion vector of the MB is such that pixel i is predicted from pixel j in the previous frame. Thus, the encoder prediction of this pixel is $\hat{f}_{n-1}^j$. The prediction error, e_n^i, is compressed, and we denote the quantized residue by $\hat{e}_n^i$. The encoder reconstruction of this pixel, $\hat{f}_n^i$, is obtained by adding the quantized residue to the prediction. Thus

$$\hat{e}_n^i = \hat{f}_n^i - \hat{f}_{n-1}^j. \tag{4}$$

What is actually transmitted over the network is the compressed residue $\hat{e}_n^i$ and the motion vector. If the current packet is correctly 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, $\tilde{f}_{n-1}^j$, which is potentially different from the value used by the encoder. Thus the decoder reconstruction of pixel i is given by

$$\tilde{f}_n^i = \hat{e}_n^i + \tilde{f}_{n-1}^j. \tag{5}$$

(Note that the value, $\tilde{f}^j_{n-1}$, is unknown to, and therefore modeled as a random variable by, the encoder.) This explains how the error propagates in time even when subsequent frames are correctly received. The probability of the packet correctly reaching the receiver is $1-p$. If the packet containing the inter-coded MB is lost, the decoder performs error concealment in a manner identical to that of an intra-coded MB. The first and second moments of $\tilde{f}^i_n$ for a pixel in an inter-coded MB are then given by (inter-mode denoted by P)

$$\begin{aligned} &E\left\{\tilde{f}^i_n\right\}(P) \\ &\quad = (1-p)\left(\hat{e}^i_n + E\left\{\tilde{f}^j_{n-1}\right\}\right) \\ &\qquad + p(1-p)E\left\{\tilde{f}^k_{n-1}\right\} + p^2 E\left\{\tilde{f}^i_{n-1}\right\} \qquad (6) \\ &E\left\{\left(\tilde{f}^i_n\right)^2\right\}(P) \\ &\quad = (1-p)E\left\{\left(\hat{e}^i_n + \tilde{f}^j_{n-1}\right)^2\right\} \\ &\qquad + p(1-p)E\left\{\left(\tilde{f}^k_{n-1}\right)^2\right\} + p^2 E\left\{\left(\tilde{f}^i_{n-1}\right)^2\right\} \\ &\quad = (1-p)\left(\left(\hat{e}^i_n\right)^2 + 2\hat{e}^i_n E\left\{\tilde{f}^j_{n-1}\right\} + E\left\{\left(\tilde{f}^j_{n-1}\right)^2\right\}\right) \\ &\qquad + p(1-p)E\left\{\left(\tilde{f}^k_{n-1}\right)^2\right\} + p^2 E\left\{\left(\tilde{f}^i_{n-1}\right)^2\right\}. \qquad (7) \end{aligned}$$

We reemphasize that these recursions are performed at the *encoder* in order to calculate the expected distortion at the *decoder*. The encoder can exploit this result directly in its encoding decisions, and, in particular, for mode switching.

If small nonlinearities such as clipping are neglected, the above derivation is precise in the case of integer-pixel motion compensation. In the half-pixel motion compensation case, we need to take into account the bilinear interpolation performed for motion compensated prediction. The first-order moment can still be computed exactly, but the second-order moment involves computing the correlation of large matrices, and appears impractical to implement in systems of reasonable complexity. However, we will show that the optimal estimate for this case is well approximated by the simpler recursion of integer-pixel motion and, although strictly speaking it is suboptimal, substantial gains are maintained. It should be noted that when coding options such as OBMC (Annex F of H.263+) or deblocking filter (Annex J of H.263+), or more sophisticated error concealment techniques are used, an appropriate modification should be made to our model.

C. Simulation Results

We now discuss the accuracy of the proposed "recursive optimal per-pixel estimate" (ROPE). We will compare our estimate to the approach recently proposed in [13] and [18], which we will refer to as the "block-weighted distortion estimate" (BWDE), and briefly specify next. BWDE estimates the decoder distortion via the formula $\hat{D} = pD_c + (1-p)D_q$. This formula should be interpreted as follows [21]: for each block of the previous frame, compute the distortion we would incur if it were lost and concealed. D_c of a block in the current frame is defined as the weighted average of the concealment distortion of the previous frame blocks that are mapped to it by motion compensation (where the weights correspond to their relative coverage of the block area) to which we add the quantization distortion of the current block residual. D_q consists only of the quantization distortion of the current residual. Note that this approach assumes that the current block is received correctly and considers two cases depending on whether or not corresponding previous frame blocks were lost and concealed. Note further that it assumes that the distortion is additive in its concealment and quantization components. As an additional reference approach, we use the simplistic estimate based only on the quantization distortion, which we call the "quantization distortion estimate" (QDE).

We implemented the distortion estimation algorithm by appropriately modifying the Telenor H.263 codec [26]. Each packet was assumed to contain one GOB and a random packet loss generator is used to drop packets at a specified loss rate. The temporal-replacement method was used for error concealment in all the cases. In the simulation, MB's are randomly selected for intra-updates. Color QCIF sequences are encoded, and the total decoder frame distortion of the *luminance* component is estimated by the above three methods. The estimates are compared to the actual decoder distortion averaged over 30 different channel realizations (with different packet loss patterns).

In Fig. 3(a) the sequence *carphone* of 250 frames is encoded with integer motion compensation at bit rate of 100 kps, frame rate of 10 f/s (a total of 84 frames encoded), and packet loss rate of 10%. In Fig. 3(b) the sequence *salesman* of 150 frames is coded at bit rate 300 kps, frame rate 30 f/s, and packet loss rate 10%. It is evident that the proposed ROPE model provides a highly accurate estimate of the decoder distortion. We also conducted similar experiments for the half pixel motion compensation case. Fig. 4 gives a typical result, and we can see that the ROPE model still provides good estimates and substantially outperforms the other competitors.

IV. RD-Based Mode Switching Algorithm

Mode switching within a rate-distortion framework is a known efficient tool for video compression in error-free channels [22], [23]. In our approach, we incorporate the overall expected distortion as computed by the ROPE model within the rate-distortion framework in order to automatically choose the *number* and the *location* of the intra-coded MB's. This enables minimization of the overall distortion (including compression and concealment) for the given packet loss rate and bit rate. We refer to the resulting technique as ROPE-RD. In this section, we consider the basic system where no feedback information is available at the encoder. The ROPE-RD algorithm is extended in Section V to the case of a communication system with a feedback channel.

A. Mode Switching Within a Rate-Distortion Framework

The "classical" rate-distortion problem in video coding is that of switching between the coding modes per MB to minimize the

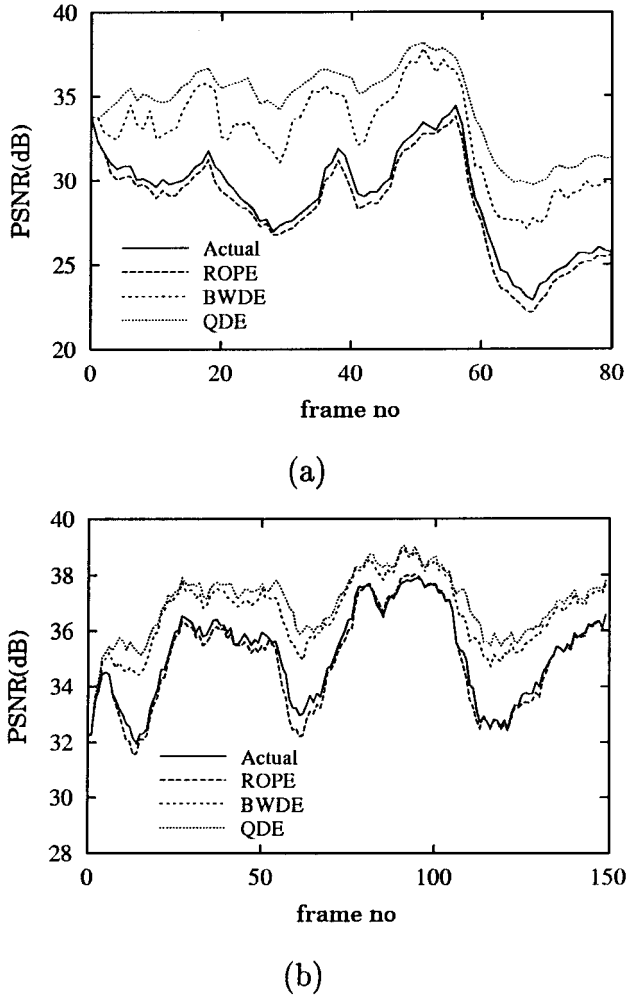


Fig. 3. Comparison between actual and estimated decoder PSNR in the integer-pixel motion compensation case. Competing estimators: ROPE (proposed), BWDE [13], [18], QDE. (a) *Carphone*, $r = 100$ kb/s, $f = 10$ f/s, $p = 10\%$. (b) *Salesman*, $r = 300$ kb/s, $f = 30$ f/s, $p = 10\%$.

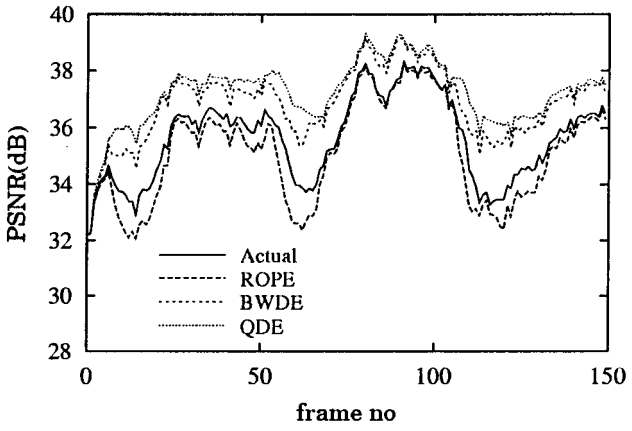


Fig. 4. Comparison between actual and estimated decoder PSNR in the half-pixel motion compensation case. Competing estimators: ROPE (proposed), BWDE [13], [18], QDE. *Salesman*, $r = 300$ kb/s, $f = 30$ f/s, $p = 10\%$.

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

$$\min_{\text{mode}}(J_{\text{MB}}) = \min_{\text{mode}}(D_{\text{MB}} + \lambda R_{\text{MB}}) \tag{8}$$

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

$$D_{\text{MB}} = \sum_{i \in \text{MB}} d_n^i. \tag{9}$$

The ROPE model is used to calculate the distortion *per pixel*, and then decide on the coding mode *per MB* via (8).

Note that this overall rate-distortion framework 1) subsumes the standard rate-distortion case [22] where packet-loss is ignored; and 2) encompasses all sources of distortion. The rate can be controlled by varying the Lagrange multiplier λ. A simple rate control scheme for rate-distortion optimized video coding in an error-free environment can be found in [22]. However, we observed that in our error-prone environment, this method doesn't work well. Instead, we performed rate-control by using the "buffer status" to update λ as in [23] and [24]. We update λ per frame via

$$\lambda_{n+1} = \lambda_n \left(1 + \alpha \left(\sum_{i=1}^{n} R_i - nR_{\text{target}}\right)\right) \tag{10}$$

where α is given by

$$\alpha = \frac{1}{5R_{\text{target}}}. \tag{11}$$

For each MB, the *mode* and the *quantization step size* are selected to minimize the rate-distortion Lagrangian. More sophisticated rate control methods are expected to achieve better performance, but this issue is peripheral to the central contribution of this paper.

The algorithm can be extended to incorporate the choice of motion vector within the rate-distortion optimization as in [25].

B. Complexity Considerations

An important aspect of the proposed ROPE-RD approach is that the expected distortion is precisely computed for each pixel. The superiority of this approach over other known methods, which calculate the distortion approximately at the MB level [13], [16], [18], will be demonstrated by the simulation results. This advantage is obtained at the cost of a modest increase in computational complexity as is explained next.

In the ROPE-RD approach, we need to compute the distortion and two moments of $\tilde{f}_n$ for the cases of intra-mode and inter-mode, for each pixel. This task forms most of the computational burden. From (1) to (7), it is easy to see that error concealment is identical regardless of the coding mode of the MB, and this fact reduces the computational complexity. For each pixel in an intra-coded MB, we need 11 addition/multiplication operations to calculate the moments of $\tilde{f}_n$. A pixel belonging to an inter-coded MB requires 16 addition/multiplication operations for this calculation. This is comparable to the number of operations needed for performing the DCT operation. (For ex-

ample, in the Telenor H.263 codec [26], the average number of addition/multiplication operations per pixel for DCT is about 24.) Further, the error concealment algorithm has to be implemented at the encoder for each block. The temporal replacement method that we use in our simulations requires negligible additional complexity. More sophisticated concealment algorithms could result in additional encoder complexity. Note that all the additional complexity is incurred *only at the encoder*.

Finally we note that we also need to store the moments as two floating-point numbers per pixel. This additional storage complexity should not pose significant difficulties in most applications.

C. Simulation Results

We implemented the ROPE-RD mode switching strategy by appropriately modifying the Telenor H.263 codec [26]. We assume the RTP payload format for packetizing the H.263 video stream [27], and that each packet contains only one GOB (other packetization formats may also be used with minor to no modification to the distortion model). A random packet loss generator is used to drop packets at a specified loss rate. The temporal-replacement method for error concealment stated in Section III-A is used in all simulations, and the rate control scheme of Section IV-A is used in all RD approaches (ROPE-RD, BWDE-RD, and QDE-RD). The color QCIF sequence is encoded at the specified bit rate by the H.263 baseline encoder. The peak signal-to-noise ratio (PSNR) of the decoder *luminance* reconstruction is computed for each frame and averaged over the whole sequence. We average the PSNR over 30 different channel realizations (with different packet loss patterns).

We first demonstrate that the accurate distortion estimate provided by the ROPE model translates into substantial gains in PSNR. We incorporated the three distortion estimation methods of Section III-C within the RD framework for mode selection. Fig. 5(a) shows the performance of the three RD optimized mode selection schemes versus packet loss rate for the *carphone* sequence encoded with integer pixel motion estimation. In Fig. 5(b), we present the results on sequence *salesman* with half pixel motion compensation. The results clearly show that precise distortion estimation enables ROPE-RD mode selection to achieve significant performance gains over the other methods.

We then compare the proposed ROPE-RD mode switching algorithm to three known mode selection schemes for packet loss networks: BWDE-RD [13], [18], "Scattered-Block Intra Update" (SB-IU), and "Contiguous-Block Intra Update (CB-IU)" [1]. BWDE-RD of [13] and [18] estimates the overall decoder distortion as explained in Section III-C [21] and incorporates the distortion into an RD framework. The SB-IU method arbitrarily assigns MB's to $1/p$ groups, and cycles through the groups updating one group per frame so that the intra-updating frequency for each MB is $1/p$ (such a method was proposed in [13] besides BWDE-RD). The CB-IU method follows the suggestion in [1], where contiguous-block patterns are recommended depending on the packet loss rate. Specifically, we use sizes of 2×2 for $p = 5\%$, 3×3 for $p = 10\%$, 4×4 for $p = 15\%$, and 5×5 for $p = 20\%$. The results under various bit rates and packet loss rates are presented.

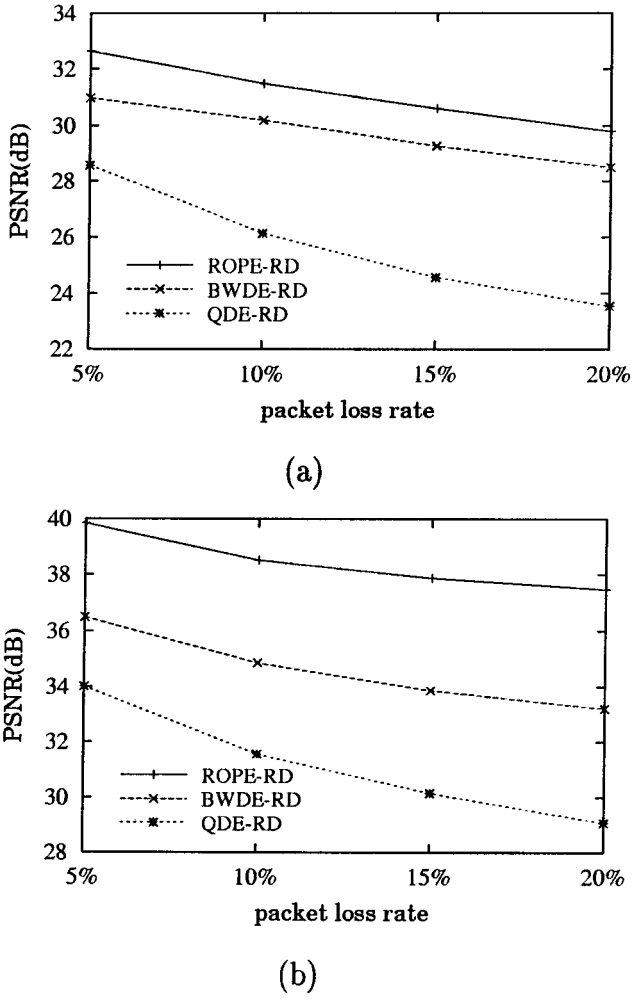


Fig. 5. PSNR versus packet loss rate for RD methods with different distortion estimates. Methods: ROPE-RD (proposed), BWDE-RD [13], [18], QDE-RD. (a) *Carphone*, integer pixel motion compensation, $r = 100$ kb/s, $f = 10$ f/s. (b) *Salesman*, half-pixel motion compensation, $r = 300$ kb/s, $f = 30$ f/s.

Tables I–III summarize simulation results for QCIF sequences *miss_america*, *grandma*, *salesman*, *mother and daughter*, *carphone*, and *foreman*. We use 150 frames in *miss_america* and 250 frames in the other sequences. Tables I and II show the results for the case of frame rate of 10 f/s, with bit rate of 64 and 100 kb/s, respectively. Results for bit rate of 300 kb/s and frame rate of 30 f/s are provided in Table III. The packet loss rate is 10% in all tables.

The tables strongly support the claim that the proposed ROPE-RD algorithm yields consistent and significant gains over the other methods. Note that the relative performance of the other approaches depends on the simulation conditions. In particular, it is noteworthy that BWDE-RD works better than SB-IU and CB-IU in sequences that exhibit substantial motion (*carphone* and *foreman*) at low frame rates (where the motion between the encoded frames is accentuated). However, this advantage diminishes and it may even become counterproductive with increased frame rate and less motion.

Fig. 6 shows the performance versus packet loss rate for the proposed ROPE-RD and the three competing methods. Results are given for the cases of integer pixel motion compensation and half-pixel motion compensation. On the *salesman*

TABLE I
PERFORMANCE COMPARISON ON QCIF SEQUENCES. (INTEGER PIXEL MOTION COMPENSATION, $r = 64$ kb/s, $f = 10$ f/s, $p = 10\%$)

Sequence	ROPE-RD	BWDE-RD	SB-IU	CB-IU
Miss America	37.78dB	37.15dB	37.56dB	35.70dB
Grandma	35.41dB	34.17dB	33.99dB	33.65dB
Salesman	33.63dB	31.56dB	31.04dB	30.94dB
Mother/Daughter	32.76dB	30.66dB	31.73dB	31.29dB
Carphone	29.89dB	28.09dB	28.02dB	27.60dB
Foreman	26.73dB	25.03dB	24.47dB	24.14dB

TABLE II
PERFORMANCE COMPARISON ON QCIF SEQUENCES. (INTEGER PIXEL MOTION COMPENSATION, $r = 100$ kb/s, $f = 10$ f/s, $p = 10\%$)

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 III
PERFORMANCE COMPARISON ON QCIF SEQUENCES. (INTEGER PIXEL MOTION COMPENSATION, $r = 300$ kb/s, $f = 30$ f/s, $p = 10\%$)

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

sequence, ROPE-RD outperforms the other three methods by 1.9∼4.8 dB in the integer-pixel motion compensation case, and by 1.5∼4.2 dB for half-pixel motion compensation case. We believe that the greater gains in the integer pixel motion compensation case are mainly due to the fact that the distortion computation is exact. Fig. 7 shows the performance versus bit rate on the *salesman* sequence for packet loss rate of 10% at frame rate 30 f/s. Fig. 8 gives the frame-by-frame evolution of the average PSNR for the *carphone* sequence when encoded at 100 kb/s and 10 f/s with packet loss rate of 10%.

It is possible to encounter mismatch between the packet loss rate assumed by the encoder in its optimization, and the actual packet loss rate in the network. Fig. 9 depicts the performance when the encoders assume $p = 5\%$ while the actual packet loss rates are 10%, 15%, or 20%. Packet loss mismatch degrades the performance of all the methods, but the proposed ROPE-RD exhibits a somewhat better robustness to mismatch, and increases its gains over the competing approaches.

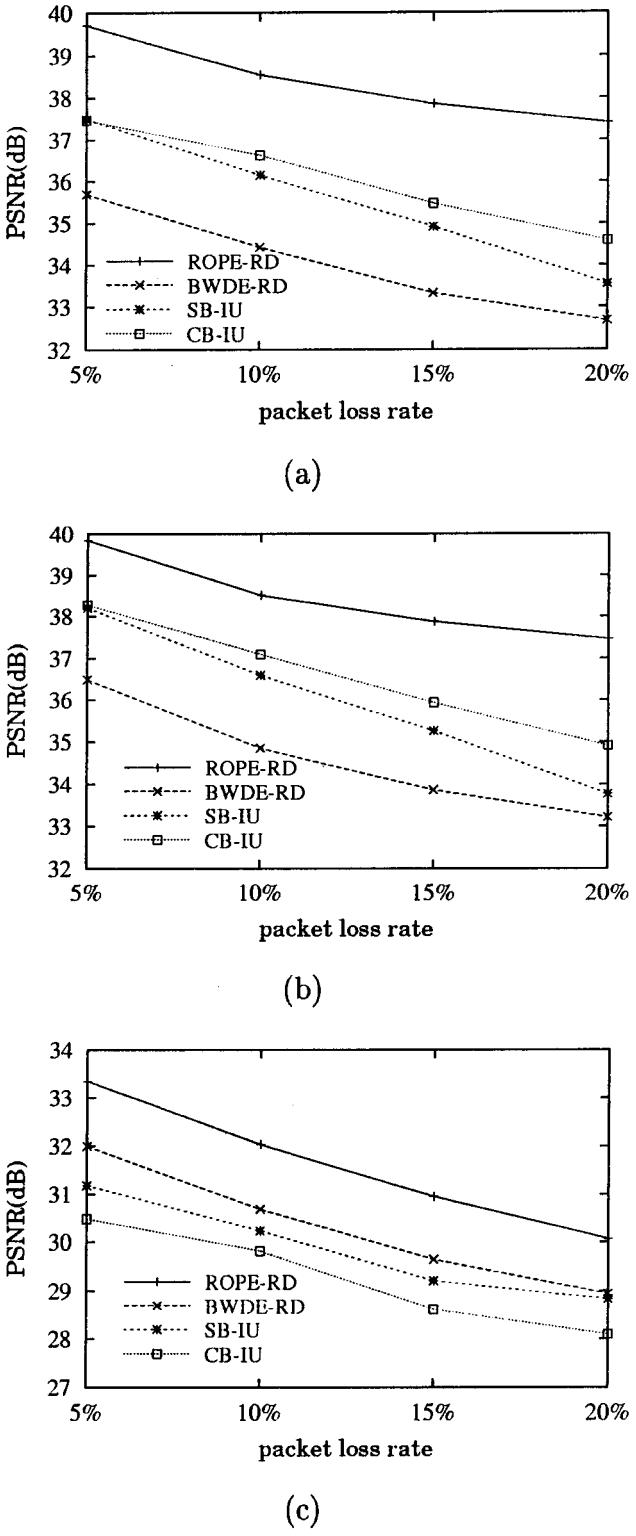


Fig. 6. PSNR versus packet loss rate. Methods: ROPE-RD (proposed), BWDE-RD [13], [18], SB-IU [13], CB-IU [1]. (a) *Salesman*, integer pixel motion compensation, $r = 300$ kb/s, $f = 30$ f/s. (b) *Salesman*, half-pixel motion compensation, $r = 300$ kb/s, $f = 30$ f/s. (c) *Carphone*, half-pixel motion compensation, $r = 100$ kb/s, $f = 10$ f/s.

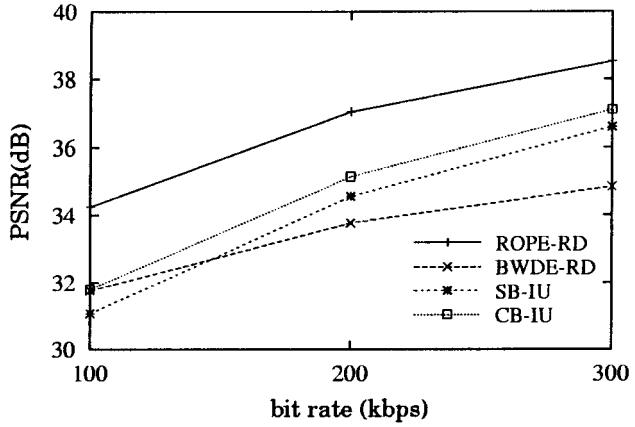


Fig. 7. PSNR versus bit rate. Methods: ROPE-RD (proposed), BWDE-RD [13], [18], SB-IU [13], CB-IU [1]. *Salesman*, half-pixel motion compensation, $f = 30$ f/s, $p = 10\%$.

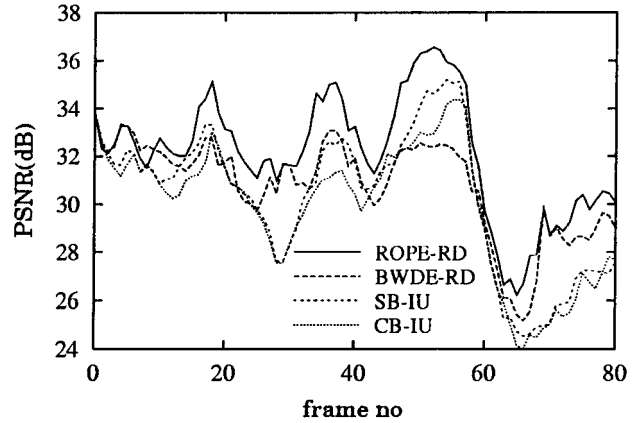


Fig. 8. PSNR versus time. Methods: ROPE-RD (proposed), BWDE-RD [13], [18], SB-IU [13], CB-IU [1]. *Carphone*, half-pixel $r = 100$ kb/s, $f = 10$ f/s, $p = 10\%$.

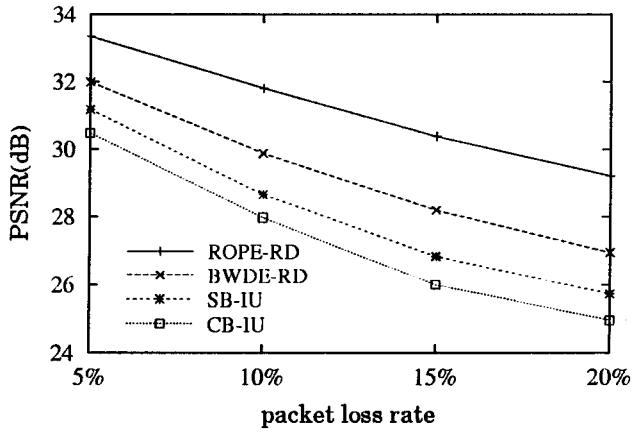


Fig. 9. Performance for mismatch in packet loss rate. Methods: ROPE-RD (proposed), BWDE-RD [13], [18], SB-IU [13], CB-IU [1]. *Carphone*, half-pixel motion compensation, assumed $p = 5\%$, $r = 100$ kb/s, $f = 10$ f/s.

V. Incorporating Information from a Feedback Channel

So far, we have assumed that there is no feedback information from the decoder. However, in certain practical applications, a backward channel from the receiver to the transmitter is available. Through this channel, the receiver can signal to the transmitter which packets were lost. We next extend the proposed algorithm to naturally incorporate such feedback information.

A. Estimate Refinement with Feedback Information

We assume that a backward channel can indicate lost packets via acknowledgment (ACK) or negative-acknowledgment (NAK). For example, RTCP [2] allows such feedback information. The feedback information is assumed to arrive error free at the encoder but with some delay. After the encoder finishes encoding the nth frame, it receives feedback information of the $(n-d)$th frame, where d is the number of frames encoded during the round trip delay. In this case, the encoder has access to the exact, albeit delayed, status of the decoder. Thus, the encoder can now exactly compute the decoder reconstruction at, and prior to, frame $(n-d)$. However, the packet loss history from frame $(n-d+1)$ to frame n remains unknown at this point. The decoder reconstruction of frames $(n-d+1)$, $(n-d+2), \cdots, n$ must still be treated as a sequence of random signals by the encoder.

We now extend the ROPE estimate to the feedback case. After getting the acknowledgment with delay d, the encoder first computes exactly the $(n-d)$th frame of decoder reconstruction, by employing error concealment whenever there was loss. Then this reconstructed frame is used to initialize the recursion formulas to compute the first and second moments of random variables through frames $(n-d+1), (n-d+2), \cdots, n$. Thus, the information obtained via feedback is utilized in refining the estimate of the overall distortion at the decoder precisely at the pixel level. Possible losses from frame $(n-d+1)$ to frame n are taken into account in the expectation. This refined estimate is then incorporated into the rate-distortion Lagrangian cost and used to optimize mode selection as is explained in the next subsection. Moreover, besides tracking the decoder status, the encoder can also adapt to variations in packet loss rate p, according to the available feedback information. This helps to track the network condition and decreases the possibility of mismatch in packet loss rate.

The computational complexity in this case is higher than that of the previous (no-feedback) case since we need to retrack the moments of the previous d frames of $\tilde{f}$ in order to refine the expected distortion of the current frame. For d in the range of $0 \sim 15$ (equivalently, 0$\sim$500 ms for 30 f/s), this complexity is still modest. Furthermore, when the round trip delay is large, we will see that mode switching may ignore the feedback information, thereby reducing complexity, with only minimal penalty in performance.

B. Simulation Results

As before, the approach is implemented by modifying the H.263 coder. We compressed 200 frames from each of the two QCIF video sequences *carphone* and *grandma*. We compare the proposed ROPE-RD method with the current state-of-the-art "error tracking" method proposed in [16] for mode switching over channels with feedback. The error tracking method intra-updates the MB's whose "error energy" is greater than the threshold. The "error energy" of the MB is initialized as

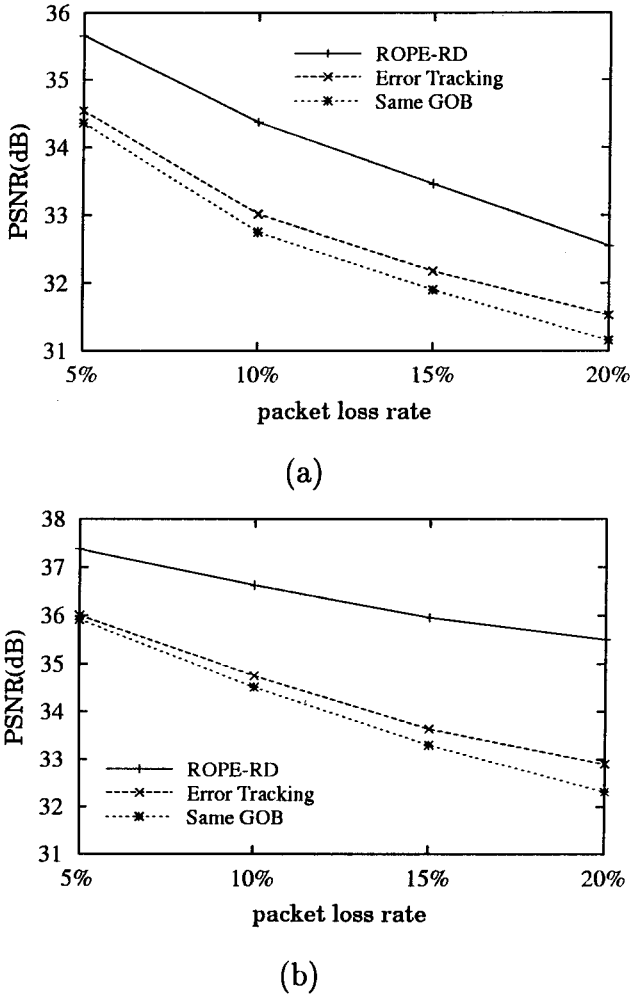


Fig. 10. PSNR versus packet loss rate for channels with a feedback delay $d = 500$ ms. Methods: ROPE-RD (proposed), Error Tracking [16], Same GOB. (a) *Carphone*, half-pixel motion compensation, $r = 300$ kb/s, $f = 30$ f/s. (b) *Grandma*, half-pixel motion compensation, $r = 300$ kb/s, $f = 30$ f/s.

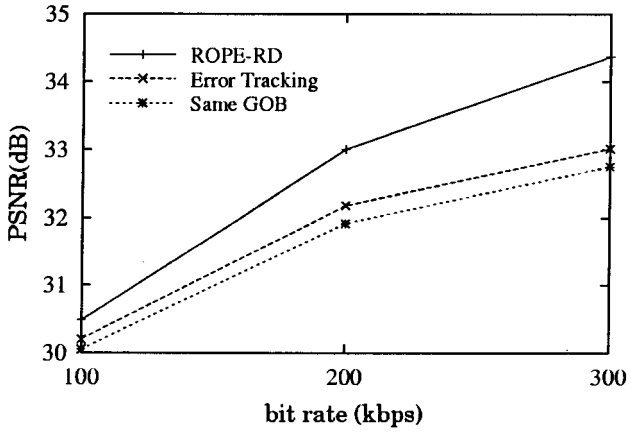


Fig. 11. PSNR versus bit rate for channels with a feedback delay $d = 500$ ms. Methods: ROPE-RD (proposed), Error Tracking [16], Same GOB. *Carphone*, half-pixel motion compensation, $f = 30$ f/s, $p = 10\%$, $d = 500$ ms.

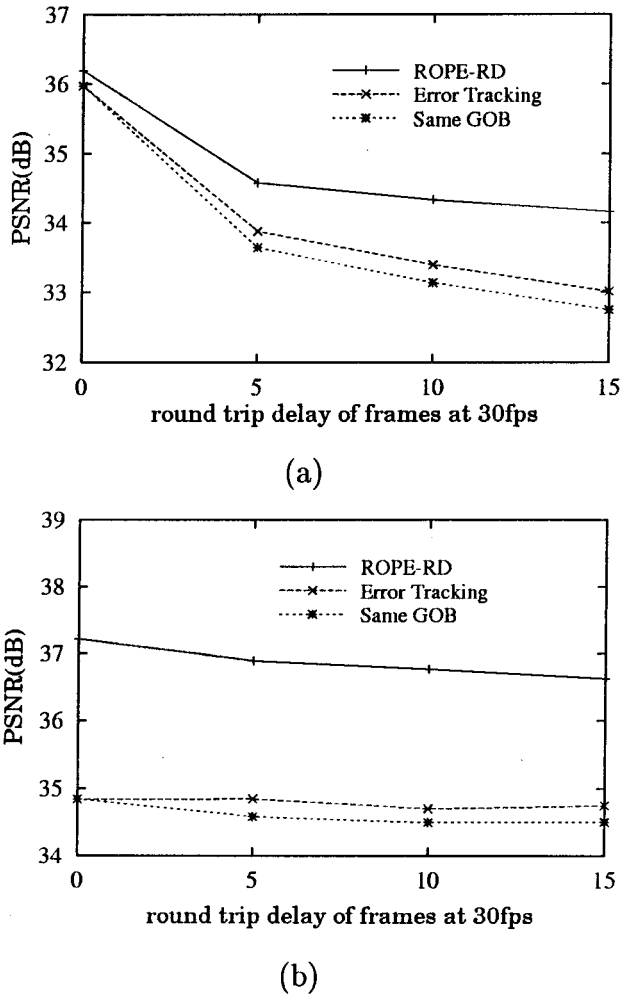


Fig. 12. PSNR versus feedback delay. Methods: ROPE-RD (proposed), Error Tracking [16], Same GOB. (a) *Carphone*, half-pixel motion compensation, $r = 300$ kb/s, $f = 30$ f/s, $p = 10\%$. (b) *Grandma*, half-pixel motion compensation, $r = 100$ kb/s, $f = 30$ f/s, $p = 10\%$.

the sum of absolute differences between the original block and the reconstructed block whenever the MB is reported as lost, and updated through temporal and spatial propagation given the motion vectors [16]. In our simulation, we use the threshold of 200. For additional reference, we provide the performance of the "same GOB" method mentioned in [16]. In this method, MB's of the current frame at the location of MB's that are reported as lost (in a prior frame) are intra-coded. The "same GOB" scheme does not consider spatial error propagation. While the "error tracking" method takes this into account, the estimate is imprecise, the updates are at the MB level, and a heuristic threshold is used. Further, both these methods ignore the effects of potential packet loss from frame $n - d + 1$ to frame n. The temporal-replacement described in Section III-A is used for error concealment by all the competing methods in the simulations. The results under various bit rates, packet loss rates, and feedback delays are presented and discussed.

Fig. 10 shows the performance versus packet loss rate for a feedback channel with a delay of 500 ms. Fig. 11 presents the performance versus bit rate on the *carphone* sequence for packet loss rate of 10% and delay of 500 ms. The figures provide ample evidence for the superiority of ROPE-RD, which outperforms the best of the other two methods by $0.3 \sim 2.6$ dB.

In Fig. 12, we present the performance versus feedback delay at packet loss rate of 10%. The frame rate is fixed at 30 f/s, and

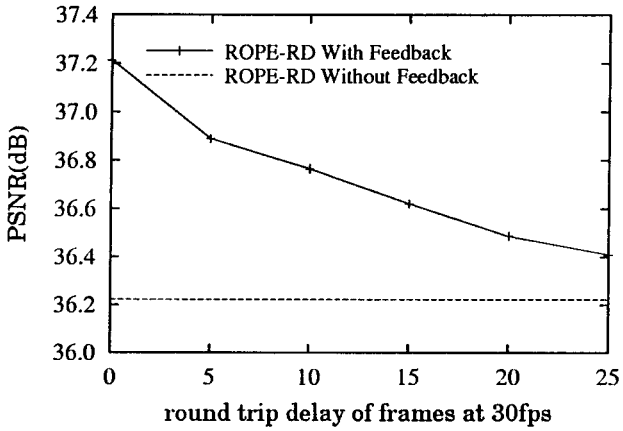


Fig. 13. Performance of ROPE-RD without feedback and with various feedback delays. *Grandma*, half-pixel motion compensation, $r = 100$ kb/s, $f = 30$ f/s, $p = 10\%$.

the delay is expressed in terms of the number of frames, $d = 0$ implies that packet loss information for the nth frame is received right after it is encoded. Note that ROPE-RD achieves larger gains as the delay increases. This is because we account for the possibility of further packet losses between the last frame, for which acknowledgment was received, and the current frame.

As is seen from Fig. 13, the performance advantage of incorporating feedback information is diminishing as the delay increases. Hence, when the delay is long enough, we can use ROPE-RD without feedback and thereby decrease the computational complexity at negligible loss in performance.

VI. Conclusion

A method is proposed for optimal intra/inter-mode switching, which enhances the robustness of video coders to packet loss. The encoder computes an optimal estimate of the total distortion of decoder frame reconstruction for the given rate, packet loss condition, and error concealment method. The algorithm recursively computes the total distortion at pixel-level precision to accurately account for both temporal and spatial error propagation. The accuracy of the estimate is demonstrated via simulation results. We incorporate the esimate within an RD framework to optimally select the coding mode for each MB. Simulation results show that our method substantially and consistently outperforms state-of-the-art RD- and non-RD-based mode switching methods, at the cost of modest additional complexity. We further extended this framework to allow for natural incorporation of feedback information from the receiver. We showed that the extended algorithm achieves considerable gains in PSNR over known mode switching techniques that use feedback. The proposed method only requires modification of the encoder decisions, and is thus standard-compatible (although, strictly speaking, such compatibility may be formally claimed only once some error concealment technique has been included in the standard).

Acknowledgment

The authors are grateful to N. Farber and G. J. Sullivan for the thorough review of, and valuable suggestions on, an early draft of this paper. They also wish to thank F. Kossentini for providing them with detailed information on the implementation of the distortion estimator in [13]. The comments of the anonymous reviewers have also been very helpful.

References

[1] Q. F. Zhu and L. Kerofsky, "Joint source coding, transport processing and error concealment for H.323-based packet video," in *Proc. SPIE, VCIP'99*, vol. 3653, San Jose, CA, Jan. 1999, pp. 52–62.

[2] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," *RFC'89*, Jan. 1996.

[3] R. Aravind, M. R. Civanlar, and A. R. Reibman, "Packet loss resilience of MPEG-2 scalable video coding algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 426–435, Oct. 1996.

[4] M. T. Orchard, Y. Wang, V. Vaishampayan, and A. R. Reibman, "Redundancy rate-distortion analysis of multiple description coding using pairwise correlating transforms," in *Proc. Int. Conf. Image Processing, ICIP'97*, CA, Oct. 1997, pp. 608–611.

[5] R. Swann and N. G. Kingsbury, "The EREC: An error resilient technique for coding variable-length blocks of data," *IEEE Trans. Image Processing*, vol. 5, pp. 656–574, Apr. 1996.

[6] Y. Wang, Q. F. Zhu, and L. Shaw, "Maximally smooth image recovery in transform coding," *IEEE Trans. Commun.*, vol. 41, pp. 1544–1551, Oct. 1993.

[7] H. Sun and W. Kwok, "Concealment of damaged block transform coded images using projections onto convex sets," *IEEE Trans. Image Processing*, vol. 4, pp. 470–477, Apr. 1995.

[8] Y. Wang and Q. F. Zhu, "Error control and concealment for video communication: A review," *Proc. IEEE*, vol. 86, pp. 974–997, May 1998.

[9] C. Horne and A. R. Reibman, "Adaptation to cell loss in a 2-layer video codec for ATM networks," in *Proc. 1993 Picture Coding Symp.*, Mar. 1993.

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

[11] S. Wenger, G. Knorr, J. Ott, and F. Kossentini, "Error resilience support in H.263+," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 867–877, Nov. 1998.

[12] T. Turletti and C. Huitema, "Videoconferencing on the Internet," *IEEE/ACM Trans. Networking*, vol. 4, no. 3, pp. 340–351, June 1996.

[13] G. Cote and F. Kossentini, "Optimal intra coding of blocks for robust video communication over the Internet," in *Image Commun.*, Sept. 1999, pp. 25–34.

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

[15] J. Y. Liao and J. D. Villasenor, "Adaptive intra update for video coding over noisy channels," in *Proc. Int. Conf. Image Processing, ICIP'96*, Switzerland, Sept. 1996, pp. 763–766.

[16] E. Steinbach, N. Farber, and B. Girod, "Standard compatible extension of H.263 for robust video transmission in mobile environments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 872–881, Dec. 1997.

[17] R. O. Hinds, T. N. Pappas, and J. S. Lim, "Joint block-based video source/channel coding for packet-switched networks," in *Proc. SPIE, VCIP'98*, vol. 3309, San Jose, CA, Jan. 1998, pp. 124–133.

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

[19] C. Hsu, A. Ortega, and M. Khansari, "Rate control for robust video transmission over wireless channels," in *Proc. SPIE, VCIP'97*, vol. 3024, San Jose, CA, Feb. 1997, pp. 1200–1211.

[20] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Mag.*, vol. 15, pp. 23–50, Nov. 1998.

[21] F. Kossentini, private communication, Sept. 1999.

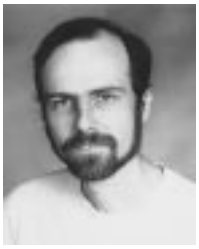
[22] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Mag.*, vol. 15, pp. 74–90, Nov. 1998.

[23] T. Wiegand, M. Lightstone, D. Mukherjee, T. George, and S. K. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 182–190, Apr. 1996.

[24] J. Choi and D. Park, "A stable feedback control of the buffer state using the controlled Lagrange multiplier method," *IEEE Trans. Image Processing*, vol. 3, pp. 546–588, Sept. 1994.
[25] R. Zhang, S. L. Regunathan, and K. Rose, "Optimal intra/inter mode switching for robust video communication over the Internet," presented at the 33rd Asilomar Conf. Signals, Syst., Computers, Oct. 1999.
[26] Telenor H.263 Codec, , ftp://bonde.nta.no/pub/tmn/software.
[27] "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.

Rui Zhang (S'00) was born in Liaoning, China, in 1973. She received the B.S. degree in electrical engineering from Beijing University of Posts and Telecommunications, China, in 1995, and the M.S. degree in electrical engineering from Tsinghua University, China, in 1997. She is now a Ph.D. student at University of California, Santa Barbara.

Shankar L. Regunathan (S'00) received the B.Tech. degree in electrical and communication engineering from the Indian Institute of Technology, Madras, in 1994, and the M.S. degree in electrical engineering from the University of California, Santa Barbara, in 1996. He is pursuing the Ph.D. degree at the University of California, Santa Barbara.



Kenneth Rose (S'85–M'91) received the B.Sc. (*summa cum laude*) and M.Sc. (*magna cum laude*) degrees in electrical engineering from Tel-Aviv University, Israel, in 1983 and 1987, respectively, and the Ph.D. degree in electrical engineering from the California Institute of Technology, in 1990.

From July 1983 to July 1988 he was employed by Tadiran Ltd, Israel, where he carried out research in the areas of image coding, image transmission through noisy channels, and general image processing. From September 1988 to December 1990 he was a graduate student at Caltech. In January 1991 he joined the Department of Electrical and Computer Engineering, University of California at Santa Barbara, where he is currently an Associate Professor. His research interests are in information theory, source and channel coding, image coding and processing, speech and general pattern recognization, and nonconvex optimization in general.

Dr. Rose is currently Editor for Source/Channel Coding for the IEEE TRANSACTIONS ON COMMUNICATIONS. He was corecipient of the William R. Bennett Prize Paper Award of the IEEE Communications Society in 1990.