

# Rate-Distortion Optimization for JVT/H.26L Video Coding in Packet Loss Environment

Thomas Stockhammer, Dimitrios Kontopodis, and Thomas Wiegand

**Abstract**—Transmission of hybrid-coded video including motion compensation and spatial prediction over error-prone channels results in the well-known problem of spatio-temporal error propagation at the decoder because of drift in the reference frames between encoder and decoder. A widely accepted and standard-compliant technique to significantly enhance the quality of the decoded video is the introduction of more intra-coded information on a macroblock basis. However, intra-coded information in general requires more bit rate. Therefore, a careful selection of intra-updates in terms of rate and distortion is necessary. A flexible and robust rate-distortion optimization technique for JVT/H.26L Coding is introduced and discussed to select coding mode and reference frame for each macroblock. The channel statistics are included in the optimization process. We derive a method to obtain an estimate of the decoder pixel distortion at the encoder. Additionally, we discuss a new method on how to choose the Lagrange parameter in packet loss environments when using a rate-control. For that, we have analytically investigated the problem. The presented techniques are verified within the new H.26L video coding standard.

## I. INTRODUCTION

Bit-streams generated by hybrid video coders including JVT/H.26L [1], [3] are extremely vulnerable to transmission errors. Transmission errors can be reduced by appropriate channel coding techniques. For channels without memory, such as the AWGN channel, channel-coding techniques provide very significant reductions of transmission errors at a comparably moderate bit-rate overhead. For the mobile fading channel and the Internet, however, the effective use of forward error correction and re-transmission is limited when assuming a small end-to-end delay. Here, the use of error resilience techniques in the source codec becomes important.

When the bit-stream is received in error, the decoder cannot or should not reconstruct the affected parts of the current frame. Rather concealment is invoked and the decoding result at the decoder differs from that at the encoder. In Inter mode, i.e., when motion-compensated prediction (MCP) is utilized, the loss of information in one frame has a considerable impact on the quality of the following frames if the concealed image content is referenced for motion compensation. As a result, spatio-temporal error propagation is a typical transmission error effect for predictive coding. Because errors remain visible for a longer period, the resulting artifacts are particularly annoying to end-users. To some extent, the impairment caused by transmission errors decays over time due to leakage in the prediction loop. However, the leakage in standardized video decoders like H.26L is not very strong, and quick recovery can only be achieved when image regions are encoded in Intra mode, i.e., without reference to a previously coded frame. The Intra mode, however, is not selected very frequently during normal encoding and completely Intra coded frames are not usually inserted in real-time encoded video as is done for storage or broadcast applications. Instead, only single macroblocks are encoded in Intra mode for regions that cannot be predicted efficiently.

The Error Tracking approach [17] utilizes the Intra mode to stop inter-frame error propagation but limits its use to severely impaired image regions only. During error-free transmission, the more effective Inter mode is utilized, and the system therefore adapts to varying channel conditions. Note that this approach

requires that the encoder has knowledge of the location and extent of erroneous image regions at the decoder. This can be achieved by utilizing a feedback channel from the receiver to the transmitter. The feedback channel is used to send NACKs back to the encoder. NACKs report the temporal and spatial location of image content that could not be decoded successfully and had to be concealed. Based on the information of a NACK, the encoder can reconstruct the resulting error distribution in the current frame, i.e., track the error from the original occurrence to the current frame. Then, the impaired macroblocks are determined and Intra coding these macroblocks can terminate error propagation. This method has the disadvantage that with an increasing round trip delay, the duration over which error propagation is visible increases and therefore only works when a small number of transmission errors occurs.

A more conservative approach is to transmit a number of Intra coded macroblocks anticipating transmission errors. In this situation, the selection of Intra coded macroblocks can be done either randomly or preferably in a certain update pattern. For example, Zhu [20] has investigated update patterns of different shape, such as 9 randomly distributed macroblocks, 1x9, or 3x3 groups of macroblocks. Although the shape of different patterns slightly influences the performance, the selection of the correct Intra percentage has a significantly higher influence. In [21] and [22], it is shown that it is advantageous to consider the image content when deciding on the frequency of Intra coding. For example, image regions that cannot be concealed very well should be refreshed more often, whereas no Intra coding is necessary for completely static background.

More recent work considers the use of Lagrangian macroblock mode decision [18] when assigning Intra macroblocks [12][13][14] with significant improvements in rate distortion performance. In [12][13][14], random reconstruction results at the decoder side are considered which depend on the statistics of the transmission errors that cause a concealment and the motion compensation that determines the inter-frame error propagation. The reconstruction quality at the decoder, i.e., the average decoding distortion, is determined by the source coding distortion, which quantifies the error between the original signal and the reconstructed signal at the encoder, and the divergence between encoder and decoder.

Lagrangian macroblock mode decision has been adopted by the ITU-T VCEG for their recommended encoder operation [9] together with the choice of the Lagrange parameter closely tied to the quantization parameter  $q$  via the formula [11]

$$\lambda_0 = \begin{cases} 0.85 \cdot q^2, & \text{for H.263} \\ 0.85 \cdot 2^{q/3}, & \text{for H.26L} \end{cases} \quad (1)$$

Note that the quantizer is two times the distance between two reproduction levels  $\Delta$ , i.e.  $q = \Delta/2$ .

In this paper, we discuss a new method on how to choose the Lagrange versus the quantization parameter in approaches like [12][13][14]. For that, we have analytically investigated the problem. Moreover, the method is tested within the new H.26L video coding standard. Based on the results presented here, our contribution to VCEG has been adopted, where in contrast to [12][13][14], we also present a new generic technique to model the average decoding result.

## II. PRELIMINARIES

### A. Notations and Formalization

Video encoding is based on a sequential encoding of frames denoted with the index  $n$ ,  $n = 1, \dots, N$  with  $N$  the total number of frames to be encoded. In most existing video coding standards including JVT/H.26L, within each frame video encoding is typically based on sequential encoding of macroblocks denoted with index  $m$ ,  $m = 1, \dots, M$  where  $M$  is total number of macroblocks in one frame and depends on the spatial resolution of the video sequence. Macro-blocks are with size  $\sqrt{I} \times \sqrt{I}$  pixel, i.e. one macroblock contains  $I$  pixel (with  $I$  being 256 in H.263 and JVT/H.26L) and the position is denoted with  $i$  where  $i = 1, \dots, I$ . The pixel value in the original sequence in frame  $n$  and macroblock  $m$  at macroblock position  $i$  is denoted as  $s_{n,m,i}$ .

We continue by formalization of the channel behavior. In the case of packet losses, the channel behavior  $c$  when transmitting frame  $n$  is defined by a binary sequence  $\{0,1\}^{\pi(n)}$  with  $\pi(n)$  the number of packets necessary to transmit frame  $1, \dots, n$ . A “0” in the channel sequence indicates a correct received packet whereas a “1” indicates a lost packet. We denote the binary channel loss sequence up to frame  $n$  as  $c_{\pi(n)}$  indicating the length of this sequence  $\pi(n)$  in the index. We can assume that the decoder is aware of the channel behavior as appropriate error detection mechanisms as block check sequences and sequence numbering are applied in common transport protocols. Additionally, let us assume that the decoder after decoding the received sequence of packets represents the pixel at position  $i$  in macroblock  $m$  and frame  $n$  with the pixel value  $\hat{s}_{n,m,i}$ . We define the distortion  $d_{n,m,i}$  by representing this pixel at the specified position as quadratic error

$$d_{n,m,i} = \left| s_{n,m,i} - \hat{s}_{n,m,i} \right|^2. \quad (2)$$

The reconstructed value  $\hat{s}_{n,m,i}$  depends on coding option selection of the encoder, the channel behavior  $c$  and, in case of error-prone channel behavior, the error concealment in the decoder. However, as common video coding schemes are hybrid and employ inter-frame prediction, the encoder includes a decoder. To avoid the problem of drift and error propagation, both encoder and decoder have to access identical reference frames with pixel values  $\hat{s}_{n,m,i}$ . Therefore, to fulfill this requirement in an error-prone environment, the encoder has to know not only the encoding options but also the channel behavior  $c$  and the error concealment applied in the decoder. Whereas the applied concealment is negotiable in a setup procedure, the channel is in general unknown at the encoder as packet losses occur randomly. We define the channel as a random variable channel denoted as  $C_{\pi(n)}$ .

### B. Problem Formulation

Although mechanisms applying feedback in the video transmission system can be used [17], the feedback is in general not instantaneous and, therefore, usually only a delayed version of the channel  $c_{\pi(n-d)}$  with  $d \geq 1$  is available at the encoder when encoding frame  $n$ . However, in the following we assume that  $d \gg 1$ , i.e. the channel is not known at the encoder. Therefore, the inherent problem of drift between reference frames in encoder and decoder cannot be avoided. Means to reduce or eliminate the drift are necessary. A very common approach is the introduction of regular I-frames especially in cases where random access to the sequence is necessary. However, in terms of bit rate and delay this is a rather unskillful solution. Therefore, to reduce the instantaneous bit-rate caused by an entire I-frame, common video standards allow introducing intra-updates on macroblock basis. By switching off the inter-frame

prediction loop for certain macroblocks the error propagation can be stopped. The obvious question is now when to insert an intra macroblock. Too many intra macroblocks will degrade the compression efficiency significantly whereas too little intra macroblocks degrade the error robustness of the video sequence. A detailed summary of available algorithms is presented in [12]. We will discuss a general rate-distortion optimized approach and show a very robust but flexible method to an RD based mode and reference frame selection.

### III. OPTIMIZED ENCODER CONTROL

#### A. RD based Macroblock and Reference Frame Selection

Hybrid video coding consists of the motion compensation and the residual coding stage. The task of residual coding is to refine signal parts that are not sufficiently well represented by motion-compensated prediction. From the viewpoint of bit allocation strategies, the various modes relate to various bit rate partitions. The concept of selecting appropriate coding options in many source-coding standards is based on rate-distortion based algorithms. The two cost terms “rate” and “distortion” are linearly combined and the mode is selected such that the total cost is minimized. This can be formalized by defining the set of selectable coding options for one macroblock as  $\mathcal{D}$ . In hybrid video coding systems the macroblock mode can be selected from the set of macroblock modes  $\mathcal{M}$ . In the following, we assume that we only transmit one I-picture at the beginning of the video sequence and P-pictures for the remainder. However, the presented algorithm can be extended easily to other picture types like B or multi-hypothesis pictures. Therefore, we assume that the set of macroblock modes consists of two subsets, one including macroblock modes, which employ temporal prediction, denoted as  $\mathcal{M}_p$  and one including pure intra coding without any prediction denoted as  $\mathcal{M}_i$ . Obviously, for I-pictures the macroblock mode can only be selected from  $\mathcal{M}_i$ . In advanced video coding schemes like H.263++ [2] and H.26L, not only the mode of the macroblock can be selected but also the reference frame from the set of accessible reference frames  $\mathcal{R}$  can be chosen [15][16]. The cardinality of set of reference frames  $|\mathcal{R}|$  specifies the maximum number of reference frames. The set of accessible coding options for P-frames is defined as all possible combinations of macroblock modes and reference frames, i.e.  $\mathcal{D} = \{\mathcal{M}_i, \mathcal{M}_p \times \mathcal{R}\}$ . Therefore, rate-constrained mode decision selects the coding option  $o_{n,m}^*$  for macroblock  $m$  in frame  $n$  such that the Lagrangian cost functional is minimized, i.e.

$$o_{n,m}^* = \arg \min_{o \in \mathcal{D}} (D_{n,m}(o) + \lambda R_{n,m}(o)). \quad (3)$$

The remaining problems in this approach are now the computation of the resulting distortion  $D_{n,m}(o)$  and the resulting rate  $R_{n,m}(o)$  when encoding with  $o$  and the proper selection of  $\lambda$ . The rate is simply obtained by encoding with mode  $o$  and in the error-free case, the encoder also can easily compute the introduced distortion

$$D_{n,m}(o) = \frac{1}{I} \sum_{i=1}^I d_{n,m,i}(o) = \frac{1}{I} \sum_{i=1}^I |s_{n,m,i} - \hat{s}_{n,m,i}(o)|^2, \quad (4)$$

where  $\hat{s}_{n,m,i}(o)$  is the reconstructed pixel value at the decoder in frame  $n$  and MB  $m$  at position  $i$  when encoding with mode  $o$ . The appropriate selection of  $\lambda$  is presented in [11]. However, in case of

error-prone transmission and resulting packet losses neither the Lagrangian multiplier selection is known nor the computation of the distortion is straightforward. Both issues will be discussed in the following.

### B. Estimate of Decoder Distortion

We will start addressing the problem how to compute the distortion  $D_{n,m}(o)$ . In the following, we will skip the dependency on the coding mode  $o$  as the computation is identical for each coding mode. We will focus on the computation of the pixel distortion  $d_{n,m,i}$  as the distortion of the macroblock is easily obtained by averaging over all pixel positions within the macroblock. As discussed previously, the pixel distortion  $d_{n,m,i}$  is not known at the encoder as it depends on the reconstructed pixel value  $\hat{s}_{n,m,i}$  and, therefore, on the random channel behavior. We emphasize this dependency by defining  $\hat{s}_{n,m,i}(C_{\pi(n)})$ . However, assuming the encoder has knowledge on the expectation values of  $C_{\pi(n)}$ . Then we can get an estimate at the encoder of the reconstructed value at the decoder, and, therefore of the distortion  $d_{n,m,i}$  as the expectation

$$d_{n,m,i} = E_{C_{\pi(n)}} \left| s_{n,m,i} - \hat{s}_{n,m,i}(C_{\pi(n)}) \right|^2, \quad (5)$$

where the expectation is over the channel  $C_{\pi(n)}$ . The computation of the expectation is based on the mean  $E\{\hat{s}_{n,m,i}\}$  and the variance  $E\{\hat{s}_{n,m,i}^2\}$  of the expected pixel value is presented in [12]. Though in [12] it is stated that the introduction of additional coding features like de-blocking filters or advanced concealment methods are possible, the extension is not straightforward to more advanced coding schemes like H.26L. The in-loop filter and the sub-pel motion accuracy require taking into account the expectation of products of pixels at different positions. Additionally, the use of advanced error concealment techniques further complicates the application of [12]. Therefore, the computational complexity and the storage requirement to keep track of all expectations are significant. In this paper, the focus is on the investigation of performance bounds. Therefore, we have chosen a different method for approximating the expected decoding distortion without attempting to provide a comparison of the complexity of the two methods, which is subject to future work.

Let us assume that we have  $K$  copies of the random variable channel behavior at the encoder, denoted as  $C_{\pi(n)}(k)$ . Additionally, assume that the set of random variables  $C_{\pi(n)}(k)$ ,  $k = 1, \dots, K$  are *identically* and *independently* distributed (iid). Then, as  $K \rightarrow \infty$ , it follows by the strong law of large numbers that

$$\frac{1}{K} \sum_{k=1}^K \left| s_{n,m,i} - \hat{s}_{n,m,i}(C_{\pi(n)}(k)) \right|^2 = E_{C_{\pi(n)}} \left| s_{n,m,i} - \hat{s}_{n,m,i}(C_{\pi(n)}) \right|^2 = d_{n,m,i}, \quad (6)$$

holds with probability 1. An interpretation of the left hand side leads to a simple solution of the previously stated problem to estimate  $d_{n,m,i}$ . In the encoder  $K$  copies of the random variable channel behavior and the decoder are operated. The reconstruction of the pixel value depends on the channel behavior  $C_{\pi(n)}(k)$  and the decoder including error concealment. The  $K$  copies of channel and decoder pairs in the encoder operate independently. Therefore, the expected distortion at the decoder can be estimated accurately in the encoder if  $K$  is chosen large enough. Obviously, the method is rather complex as  $K$  times the complexity and the memory requirement of decoder is necessary in the encoder. However, due to the simplicity, robustness, and flexibility of the approach and the good converging properties for even low  $K$  this approach is very suitable to obtain performance bounds.

### C. Lagrange Multiplier Selection in Packet-Loss Environment

The selection of the Lagrangian multiplier for the macroblock mode for error-free transmission has been discussed in details in [9] and [11]. It is known that the Lagrangian multiplier corresponds to the negative slope of the distortion-rate function. We will use similar arguments as in [9] and [11] to derive an appropriate selection for  $\lambda$ .

Additionally, it is simple to show that if the distortion-rate function  $D(R)$  is strictly convex then the functional  $J(R) = D(R) + \lambda R$  is also strictly convex. Assuming  $D(R)$  to be differentiable everywhere, the minimum of the functional  $J(R)$  can be derived by the setting the derivative to zero, i.e.

$$\lambda = -\frac{dD(R)}{dR}. \quad (7)$$

Therefore, to derive an appropriate  $\lambda$  an expression of the distortion  $D(R)$  is necessary. If we assume high-resolution quantization it is well known [23] that the source distortion  $D_s(R)$  depends on the rate as

$$D_s(R) = \beta \cdot 2^{-\alpha R}, \quad (8)$$

with  $\beta$  being a constant depending on the variance of the source. However, in contrast to the error-free case in addition to the source distortion we also have to consider the distortion incurred if the source data is lost. Obviously, this distortion depends on the error concealment in general. We will provide some arguments on how to select the Lagrange multiplier in error-prone environment. The expected overall distortion  $D(R)$  can be estimated as

$$D(R) = (1-p)p_c D_s(R) + pD^{(ec)} + (1-p)(1-p_c)D^{(ep)} \quad (9)$$

with  $p$  the loss probability of the current macroblock,  $p_c$  the probability that the referenced image part is correct,  $D^{(ec)}$  the error concealment distortion if this macroblock is lost, and  $D^{(ep)}$  the expected error propagation (ep) distortion in the case that this macroblock is received correctly but the reference frames are erroneous.

Similar to [11] it is assumed for the distortion-to-quantizer relation that at sufficiently high rates, the source probability distribution can be approximated as uniform within each quantization interval [23] yielding

$$D(\Delta) = (1-p)p_c \frac{\Delta^2}{12} + pD^{(ec)} + (1-p)(1-p_c)D^{(ep)} \quad (10)$$

with  $\Delta$  being the quantizer step size, or, equivalently the distance of the quantizer reproduction levels. Therefore, with (6), (7) and (8) we can solve for  $R(\Delta)$  as

$$R(\Delta) = \frac{1}{\alpha} \log_2 \left( \frac{\beta}{D_s(\Delta)} \right) = \frac{1}{\alpha} \log_2 \left( \frac{\beta}{\Delta^2/12} \right). \quad (11)$$

By combining the derivatives for  $\Delta$  in (8) and (9) in (5) we obtain

$$\lambda = -\frac{dD}{d\Delta} \frac{d\Delta}{dR} = \frac{\alpha \log(2)}{12} \Delta^2 (1-p) p_c = (1-p) p_c \lambda_0, \quad (12)$$

with  $\lambda_0$  the Lagrange multiplier for error-free transmission according to (1). Although the assumptions may not be completely realistic, the derivation yields some insight in the selection of the Lagrange parameter in error-prone transmission environment. Additionally, the estimation of  $p_c$  is not straightforward. However, in general the probability  $p_c$  that a certain image part is correct, decreases with increasing distance to a intra refresh of a certain region and also depends on the macroblock loss rate  $p$ . Therefore, with  $p$  increasing  $\lambda$  decreases and, therefore, the rate has less weight in mode selection.

#### D. Implementation of Mode Selection Algorithm

We will now combine the results from the previous three subsections to obtain a simple and straightforward decoder distortion estimation and mode selection in the encoder. Therefore, let us have a closer look at the expected macroblock distortion  $D_{n,m}(o)$ . According to (9)  $D_{n,m}(o)$  consists of three parts. Let us denote the source distortion  $D_s$  as error-free (ef) distortion  $D_{n,m}^{(ef)}(o)$ , the error concealment distortion as  $D_{n,m}^{(ec)}(o)$  and the error propagation distortion as  $D_{n,m}^{(ep)}(o)$ . Whereas  $D_{n,m}^{(ef)}(o)$  and  $D_{n,m}^{(ep)}(o)$  depend on the coding option  $o$ , the error concealment distortion is independent of the coding option, i.e.  $D_{n,m}^{(ec)}(o) = D_{n,m}^{(ec)}$ . Therefore, the mode selection according to (3) can be reformulated as

$$\begin{aligned} o_{n,m}^* &= \arg \min_{o \in \mathcal{D}} \left( (1-p) p_c D_{n,m}^{(ef)}(o) + (1-p)(1-p_c) D_{n,m}^{(ep)}(o) + p D_{n,m}^{(ec)} + \lambda_0 (1-p) p_c R_{n,m}(o) \right) \quad (13) \\ &= \arg \min_{o \in \mathcal{D}} \left( (1-p) p_c D_{n,m}^{(ef)}(o) + (1-p)(1-p_c) D_{n,m}^{(ep)}(o) + \lambda_0 (1-p) p_c R_{n,m}(o) \right) \\ &= \arg \min_{o \in \mathcal{D}} \left( p_c D_{n,m}^{(ef)}(o) + (1-p_c) D_{n,m}^{(ep)}(o) + \lambda_0 p_c R_{n,m}(o) \right) \\ &= \arg \min_{o \in \mathcal{D}} \left( \hat{D}_{n,m}(o) + \lambda_0 p_c R_{n,m}(o) \right). \end{aligned}$$

with  $\hat{D}_{n,m}(o) = p_c D_{n,m}^{(ef)}(o) + (1-p_c) D_{n,m}^{(ep)}(o)$  the expected distortion of the macroblock  $m$  in frame  $n$  encoded with mode  $o$ .  $\hat{D}_{n,m}(o)$  is computed as the average over the  $K$  distortions by decoding this macroblock based on the erroneous reference frames  $n - |\mathfrak{R}|, \dots, n-1$  in each decoder  $k = 1, \dots, K$ . Obviously, this distortion strongly depends on the coding mode  $o$  as for example an intra mode stops the error propagation in this macroblock. After encoding the entire frame  $n$  the reference frame buffer in each decoder  $k = 1, \dots, K$  is updated by decoding this frame  $n$  based on the channel statistics  $C_{\pi(n)}(k)$ . A remaining problem is an appropriate selection of  $p_c$ . In our implementation, we apply a very simple model. We assume that the motion vectors for all cases are zero and that choosing an intra mode always stops error propagation. Therefore, assuming that this macroblock has been coded in Inter mode for the past  $r$  reference frames and in Intra mode in frame  $n-r-1$  we obtain as estimate for  $p_c = (1-p)^r$ . This means, that with increasing  $r$  and increasing loss probability  $p$  the probability  $p_c$  that the referenced image part is correct, decreases. Better models are subject of future work.

## IV. EXPERIMENTAL RESULTS

### A. Simulation Conditions and Evaluation Criteria

Several experiments have been carried out to show the performance of different intra update strategies as well as the performance of H.26L in packet lossy environments. The following evaluation criteria are defined to illustrate the performance of the different strategies. The average distortion in frame  $n$  in channel-decoder pair  $k$  is defined as

$$D_n(k) = \frac{1}{MI} \sum_{m=1}^M \sum_{i=1}^I \left( s_{n,m,i} - \hat{s}_{n,m,i} \left( C_{\pi(n)}(k) \right) \right)^2 \quad (14)$$

and the corresponding PSNR for frame  $n$  in channel-decoder pair  $k$  is defined as

$$Q_n(k) = 10 \log_{10} \left( \frac{255^2}{D_n(k)} \right). \quad (15)$$

The experimental average of the PSNR averaged over  $K$  decoders is defined as

$$\mu_Q^{(K)}(n) = \frac{1}{K} \sum_{k=1}^K Q_n(k), \quad (16)$$

and the statistical expectation is defined as  $\mu_Q(n) = \lim_{K \rightarrow \infty} \mu_Q^{(K)}(n)$ . The experimental standard deviation of the decoder quality assuming to have  $K$  decoders is defined as

$$\sigma_Q^{(K)}(n) = \sqrt{\sum_{k=1}^K \left( Q_n(k) - \mu_Q(n) \right)^2} \quad (17)$$

and the statistical standard deviation is defined as  $\sigma_Q(n) = \lim_{K \rightarrow \infty} \sigma_Q^{(K)}(n)$ . This value expresses the variance of the estimate in the encoder if only one decoder is operated. However, if  $K'$  decoders are run in the encoder the standard deviation decreases with  $1/\sqrt{K'}$ . We define the standard deviation in case of  $K'$  decoders in the encoder as  $\sigma_{Q,K'}(n) = \sigma_Q(n) / \sqrt{K'}$ .

For all experiments the test sequence ‘‘foreman’’ (300 frames, 30 frames per second, QCIF) is encoded at 7.5 frames per second with the H.26L test model encoder JM1.4. Each row of  $16 \times 16$  macroblocks is transmitted in a separate packet, i.e. 11 macroblocks per packet. Packets are assumed to be lost independently with loss rate  $p = 0.1$ . In the encoder in complete  $K = 500$  pairs of channel and decoder are operated. It is assumed that this results in sufficient statistical significance and we assume that the statistical expectations are identical to the experimental averages, i.e.  $\mu_Q(n) = \mu_Q^{(K=500)}(n)$  and  $\sigma_Q(n) = \sigma_Q^{(K=500)}(n)$ . The quantization parameter for all experiments is chosen to  $q = 20$  unless state otherwise. The optimized macroblock mode and reference frame selection is compared to two different strategies intra update strategies. Intra updates can be selected randomly or regularly whereby the number of forced intra-updates  $\eta$  per frame can be specified. In the random mode  $\eta$  randomly chosen positions in one frame are intra updated. For each frame the intra update position are randomly picked independent for each frame. In the regular mode the  $\eta$  intra update positions are chosen sequentially. In the

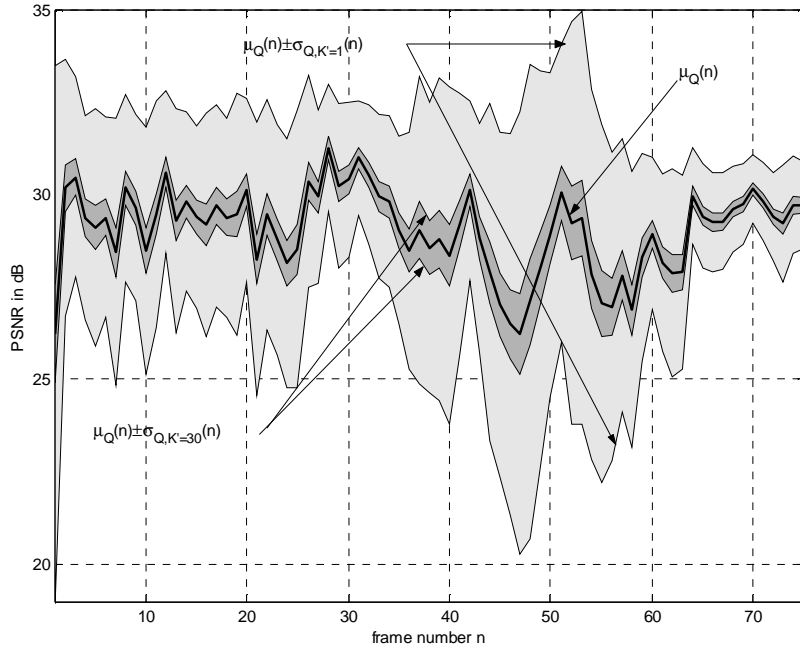
first frame the macroblock positions  $m = 1, \dots, \eta$ , in the second frame the macroblock positions  $m = \eta + 1, \dots, 2\eta$ , and so on, are intra updated. If all macroblocks have been updated once, the first positions are intra updated again. The starting position is shifted such that each frame has  $\eta$  intra updates. Therefore, for some positions macroblocks in the lower right corner as well in the upper left corner are intra-updated. Both intra update modes are combined with the RD optimized mode selection and reference frame selection based on the error-free reference frames according to [9]. As this mode selection also might choose an intra update the actual number of intra updates per frame might be higher than  $\eta$ . In addition, in case of multiple reference frames a restriction on the reference frames in combination with error-resilient intra macroblock updates can be applied. The set of selectable reference frames in the rate-distortion optimized reference frame and macroblock selection is restricted such that no pixels are used for prediction which have been intra refreshed for error resilience in later frames. This restriction can be used in combination with any intra update mode. For details of this algorithm, we refer to [4]. The error concealment in  $K$  decoders in the encoder is simple previous frame concealment. An extension to more advanced error concealment strategies [10] not only in the test model decoder but also in the pixel distortion estimation in the encoder remains work to be conducted.

### B. Decoder Distortion Estimation in Encoder

In the first experiment, we show the influence of the number of decoders  $K$  in the encoder on the estimated decoder distortion. Therefore, the test sequence “foreman” was encoded according to the conditions presented in subsection IV.A. The optimized macroblock mode selection according to subsection III.D has been applied and the number of reference frames is limited to 1 for this experiment. Figure 1 shows for each encoded frame  $n$  the average PSNR  $\mu_Q(n) = \mu_Q^{(K=500)}(n)$  and standard deviation  $\mu_Q(n) \pm \sigma_{Q,K'}(n)$  for  $K' = 1$  and  $K' = 30$  channel and decoder pairs. Although the quantization parameter is constant, the average PSNR varies over the frame number. This is as losses have different effects on different sequence parts. The standard deviation for  $K' = 1$  allows judging the variance of the quality at the receiver. Not only the average PSNR should be high but also the variance should be low. Especially in the area with the camera pan (frame number 45 to 55), losses can have significant influence on the distortion and the variance. However, it is also obvious that a quick recovery from losses is obtained due to the optimized intra update. In addition, the variance of the results can be reduced with advanced error concealment. For  $K' = 30$  we obtain an estimate on the quality and the variation on the quality in the decoder in case of operating  $K' = 30$  decoders. It can be seen that using only 30 decoders reduces the variance of the decoder quality estimate at the encoder significantly. The standard deviation is reduced significantly to about 0.5 dB for almost all frames. The maximum standard deviation is about 1 dB. Therefore, in general it is good enough to operate about  $K' = 30$  decoders in the encoder.

### C. Comparison of Encoder Strategies

We will now compare different intra coding strategies, the influence of multiple reference frames and the selection of the Lagrange multiplier. The simulation conditions are according to subsection IV.A. The average PSNR  $1/N \sum_{n=1}^N \mu_Q(n)$  as well as the bit-rate for different settings is shown in Table 1. The bit-rate is computed by only taking into account the slice header. The packet overhead (e.g. IP headers, etc.) is ignored. However, as for all experiments the same number of packets  $\pi(N) = 675$  is transmitted, the bit-rates and results are comparable.



**Figure 1** Average PSNR  $\mu_Q(n) = \mu_Q^{(K=500)}(n)$  and standard deviation  $\mu_Q(n) \pm \sigma_{Q,K'}(n)$  for  $K' = 1$  and  $K' = 30$  channel and decoder pairs over frame number  $n$  for simulation conditions according to subsection IV.A.

Let us first compare different intra update modes. For regular and random intra update modes, the intra-update parameter  $\eta$  is chosen such that the bit-rate is comparable to the optimized case. For all settings, it is obvious that the random intra-update is inferior to the regular update. This is as in the random update some image parts might not be refreshed for a long period. The optimized intra update, however, is significantly better than any regular or random update. For the random and regular intra update, it is beneficial to introduce multiple reference frames but restrict the selection of reference frames according to [4]. In this case, gains up to 1 dB in average PSNR are visible (see experiment Reg3) compared to only 1 reference frame. The non-restricted mode performs significantly worse as the error propagation is not necessarily stopped even with a forced intra update. We will now focus on the optimized mode. For the optimized mode, the restriction of reference frames is almost negligible as the mode and reference selection inherently chooses the appropriate mode (compare experiment Opt5 and Opt 6). Compared to only one reference frame (Opt3) the bit-rate when using 5 reference frames can be reduced by about 3%. These relatively low gains are obvious as reference frames further in the past are not chosen if an intra-update has been introduced due to the possible error-propagation. Intra-updates for the “foreman” sequence at 10% error rate occur quite frequently. Finally, we look at the selection of the Lagrange parameter. The differences are not very significant. With the Lagrange parameter according to (13) the rate is slightly increased for the same QP but the quality is also increased. For linear interpolation of experiments Opt1 and Opt2 in the PSNR-rate plane, we obtain slightly better result by experiment Opt3 with using the adapted Lagrange parameter. For 5 reference frames it can be seen that a bit-rate increase of about 5% results in an average PSNR gain of 0.17 dB. Although some justification is provided, that changing the Lagrange parameter in error-prone environment might be beneficial, the gains are very insignificant. Therefore, unless significantly different results with a better model on the error propagation

can be obtained, it is reasonable to use the same Lagrange parameter for error-prone transmission as used in the mode selection process in the error-free case.

**Table 1** Average PSNR and bit-rate for different intra-update modes, different number of reference frames, different quantization parameters and different restriction modes according to test conditions in subsection IV.A.

Experiment	Intra-update mode	Reference frames	Lagrange parameter	QP	Restriction	Av. PSNR	Bit-rate
Ran1	random $\eta = 46$	1	-	20	-	27.09 dB	103.94 kbit/s
Ran2		5	-	20	Off	26.58 dB	103.31 kbit/s
Ran3		5	-	20	On	27.39 dB	104.35 kbit/s
Reg1	regular $\eta = 50$	1	-	20	-	27.46 dB	102.92 kbit/s
Reg2		5	-	20	Off	27.25 dB	102.19 kbit/s
Reg3		5	-	20	On	28.37 dB	103.23 kbit/s
Opt1	optimized $K = 500$ $p = 0.1$	1	(13)	21	-	28.61 dB	92.52 kbit/s
Opt2		1	(13)	20	-	29.04 dB	104.97 kbit/s
Opt3		1	(1)	20	-	28.90 dB	101.30 kbit/s
Opt4		5	(13)	20	On	29.05 dB	103.74 kbit/s
Opt5		5	(1)	20	Off	28.88 dB	99.89 kbit/s
Opt6		5	(1)	20	On	28.88 dB	98.55 kbit/s

Additional results on the performance of H.26L when using the optimized mode selection based on  $\bar{K}$  decoder in the encoder have been investigated for standardized test conditions for RTP/IP over fixed Internet [5][6] and for RTP/IP over mobile in [7] and [8].

## V. CONCLUSIONS AND OUTLOOK

A flexible and robust rate-distortion optimization technique was introduced and discussed to select coding mode and reference frame for each macroblock. The channel statistics are included in the optimization process. We have presented a method to obtain an estimate of the decoder pixel distortion at the encoder. Additionally, we have discussed a new method on how to choose the Lagrange parameter in packet loss environment when using a rate-control. For that, we have analytically investigated the problem. The presented techniques have been verified within the new H.26L video coding standard. Future work includes the combination of the work with advanced error concealment strategies [10], combination with appropriate rate control schemes, a comparison with complexity-reduced algorithms to estimate the decoder distortion in the encoder, and, finally, the inclusions of feedback information in the mode selection.

## REFERENCES

- [1] T. Wiegand (ed.), "Working Draft Number 1, Revision 0 (WD-1)," Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-A003, January 2002.
- [2] ITU-T, Video Coding for Low Bit-Rate Communication, ITU-T Recommendation H.263, Version 1: November 1995, Version 2: January 1998, Version 3: Nov. 2000.

- [3] G. Sullivan, T. Wiegand, and T. Stockhammer, "Using the Draft H.26L Video Coding Standard for Mobile Applications," in Proc. ICIP 2001, Thessaloniki, Greece, October 2001.
- [4] T. Stockhammer and D. Kontopodis, "Error Robust Macroblock Mode and Reference Frame Restriction" Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-B102, January 2002.
- [5] S. Wenger, "H.26L over IP: The IP-Network Adaptation Layer", Proc. Packet Video Workshop 2002, Pittsburgh, PY, April 2002.
- [6] S. Wenger, "H.26L over IP", in preparation for IEEE CSVT, Special Issue on H.26L/JVT, April 2002.
- [7] T. Stockhammer, T. Oelbaum, and T. Wiegand, "H.26L/JVT Video Transmission in 3G Wireless Environments", accepted for 3G Wireless, San Francisco, CA, May 2002.
- [8] T. Stockhammer, M.M. Hannuksela, and T. Wiegand, "JVT/H.26L in 3G Wireless Environments", in preparation for IEEE CSVT, Special Issue on H.26L/JVT, April 2002.
- [9] 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.
- [10] V. Varsa, M. Hannuksela, and Y. Wang, "Non-normative error concealment algorithms," ITU-T VCEG-N62, VCEG (SG16/Q6), Fourteenth Meeting, Santa Barbara, CA, September 2001.
- [11] T. Wiegand, and B. Girod, "Lagrangian Multiplier Selection in Hybrid Video Coder Control," Proc. ICIP 2001, Thessaloniki, Greece, October 2001.
- [12] R. Zhang, S. L. Regunathan, and K. Rose, "Video Coding with Optimal Inter/Intra-Mode Switching for Packet Loss Resilience," in IEEE JSAC, vol. 18, no. 6, pp. 966-976.
- [13] G. Cote, S. Shirani, F. Kossentini, "Optimal mode selection and synchronization for robust video communications over error-prone networks," in IEEE JSAC, vol. 18, no. 6, pp. 952-965.
- [14] T. Wiegand, N. Färber, K. Stuhlmüller, and B. Girod, "Error-resilient video transmission using long-term memory motion-compensated prediction", IEEE JSAC, vol. 18, no. 6, pp. 1050-1062.
- [15] T. Wiegand and B. Girod, "Multi-frame Motion-Compensated Prediction for Video Transmission," Kluwer Academic Publishers, Sept. 2001.
- [16] T. Wiegand, X. Zhang, and B. Girod, "Long-Term Memory Motion-Compensated Prediction," IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, no. 1, pp. 70-84, Feb. 1999.
- [17] N. Färber, E. Steinbach, and B. Girod, "Robust H.263 Compatible Video Transmission Over Wireless Channels," In Proceedings of the Picture Coding Symposium, pages 575-578, 1996.
- [18] T. Wiegand, M. Lightstone, D. Mukherjee, T. G. Campbell, and S. K. Mitra. "Rate-Distortion Optimized Mode Selection for Very Low Bit Rate Video Coding and the Emerging H.263 Standard," IEEE Transactions on Circuits and Systems for Video Technology, 6(2):182-190, April 1996.
- [19] N. Färber, K. W. Stuhlmüller, and B. Girod, "Analysis of Error Propagation in Hybrid Video Coding with Application to Error Resilience," In Proceedings of the IEEE International Conference on Image Processing, volume 2, pages 550-554, Kobe, Japan, October 1999.
- [20] Q. F. Zhu and L. Kerofsky, "Joint Source Coding, Transport Processing, and Error Concealment for H.323-Based Packet Video," In Proceedings of the SPIE Conference on Visual Communications and Image Processing, volume 3653, pages 52-62, San Jose, CA, USA, January 1999.
- [21] P. Haskell and D. Messerschmitt, "Resynchronization of Motion-Compensated Video Affected by ATM Cell Loss," In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume 3, pages 545-548, 1992.
- [22] J. Liao and J. Villasenor, "Adaptive Intra Update for Video Coding over Noisy Channels," In Proceedings of the IEEE International Conference on Image Processing, volume 3, pages 763-766, Lausanne, Switzerland, October 1996.
- [23] H. Gish and J. N. Pierce, "Asymptotically Efficient Quantizing," IEEE Transactions on Information Theory, vol. 14, pp. 676-683, September 1968.