Modelling of Distortion Caused by Packet Losses in Video Transport

Yao Wang*, Zhenyu Wu†, Jill Boyce† and Xiaoan Lu*
*Polytechnic University, Brooklyn, NY 11201, USA. Email:{yao,xlu}@vision.poly.edu
†Corporate Research, Thomson Inc., Princeton, NJ 08540, USA. Email: {jill.boyce,zhenyu.wu}@thomson.net

***Abstract*— This paper analyzes transmission-error induced distortion in decoded video. A recursion model is derived that relates the distortion in successive P-frames. The model takes into account of non-integer motion vectors used for motion-compensated temporal prediction and concealment, unconstrained intra prediction, and in-loop deblocking filtering. Experimental data show that the model is quite accurate over a large range of packet loss rates and encoder intra rates.**

I. Introduction

The quality of the received video in a networked video application depends both on the quantization incurred at the encoder, the channel errors occurred during transmission, and consequent error-propagation in the decoded sequence. The channel-induced distortion depends both on channel loss characteristics and the coder error resilience features, most notably the intra-block rate at the encoder. Accurate modeling of the channel-induced distortion is important for jointly determining parameters for source coding (e.g. quantization and intra-rate) and channel error control (e.g. channel code rate, retransmission limit), and for rate-distortion optimized mode decision in the encoder. There have been several important contributions in this area, including [1], [2], [3].

All the prior works consider error propagation due to temporal prediction only and most of them do not take into account of non-integer motion compensation and deblocking filtering. Intra-prediction and deblocking filtering are two new features of the latest H.264 video coding standard and contribute significantly to the improvement of coding efficiency over prior standards. In this paper, we develop a model for channel-induced distortion that considers both inter- and intra-prediction and deblocking. The model also takes into account of decoder temporal concealment. For motion compensated prediction and concealment, we explicitly consider the effect of non-integer motion vectors.

II. Notation and Assumptions

Let f_n^i denote the original pixel value in frame n and pixel i, $\hat{f}_n^i$ the reconstructed signal at the encoder, and $\tilde{f}_n^i$ the reconstructed signal at the decoder. The channel induced distortion at pixel i is defined as $D_{c,n}^i = E_c\{(\hat{f}_n^i - \tilde{f}_n^i)^2\}$, where $E_c\{\cdot\}$ represents the expectation taken over all possible channel realizations. In this paper, we are interested in modelling the average channel-induced distortion in each frame, defined by $D_{c,n} = E_a\{D_{c,n}^i\}$, where $E_a\{\cdot\}$ denotes the averaging-over-pixel operation. In our derivation, we assume $D_{c,n}^i$ is pixel-location independent and equal to the average channel distortion in that frame, i.e., $D_{c,n} = D_{c,n}^i$. We further use $E\{\cdot\}$ to denote the concatenated operation $E_a\{E_c\{\cdot\}\}$.

We assume that the MBs in a frame are independently coded into slices, and each slice has its own header and is carried in a separate packet. We further assume that the loss of any bits in a slice will make the entire slice undecodable. We also assume that with proper packet interleaving, the packet (and hence slice) loss event can be characterized as an i.i.d. random process by a loss rate P.

We assume a video sequence is partitioned into groups of frames (GoFs) and each GoF starts with an I-frame, followed by P-frames. We consider how to model the progression of the channel distortion in successive P-frames. Within each P-frame an MB may be coded in either inter (P) or intra (I) mode. The I-mode is used either because it achieves a better rate-distortion trade-off, or for error-resilience purpose. We use β_n to denote the percentage of MBs that are coded in the I-mode in frame n. We assume that if an MB is lost in frame n, it will be concealed using motion-compensated temporal concealment, with an average distortion $D_{\mathrm{L},n}$. If an MB is received, it could still have channel distortion due to errors in previous frames or pixels in the same frame. Denoting the average distortion in received I-MBs and P-MBs by $D_{\mathrm{IR},n}$ and $D_{\mathrm{PR},n}$, respectively, the average channel distortion is

$$D_{c,n} = (1-P)((1-\beta_n)D_{\mathrm{PR},n} + \beta_n D_{\mathrm{IR},n}) + PD_{\mathrm{L},n}. \quad (1)$$

In the following section, we derive the recursion formula that relates $D_{c,n}$ with $D_{c,n-1}$ for P-frames.

III. The Distortion Model

Case I: Motion-Compensated Temporal Prediction and Concealment with Non-Integer Motion Vectors

For ease of understanding, we first develop the recursion assuming the encoder does not use intra-prediction and deblocking filtering. In this case, for a received I-MB, there will be no channel distortion, i.e., $D_{\mathrm{IR},n} = 0$. For a P-MB, even if it is received, its reconstruction may have channel distortion due to errors in the previous frame. To take into account of the interpolation operation typically applied when doing motion compensation using non-integer motion vectors, we assume a pixel f_n^i is predicted by a weighted sum of several neighboring pixels in frame n-1, denoted by $f_{p,p,e}^i = \sum_{l=1}^{L_{p,p}} a_l \hat{f}_{n-1}^{u_l(i)}$, where $u_l(i)$ refers to the spatial index of the l-th pixel in

0-7803-9332-5/05/$20.00 ©2005 IEEE

frame $n-1$ that was used to predict f_n^i. The values for $L_{p,p}$ and a_l depend on the MV for the MB, and the interpolation filter employed for fractional-pel motion compensation, with $\sum_l a_l = 1$.

In the receiver, the prediction is based on $f_{p,p,d}^i = \sum_{l=1}^{L_{p,p}} a_l \tilde{f}_{n-1}^{u_l(i)}$. For the pixels associated with a particular set of $L_{p,p}, a_l$, the channel distortion is, with $e_{n-1}^i = \hat{f}_{n-1}^i - \tilde{f}_{n-1}^i$,

$$\begin{aligned} D_{\mathrm{PR},n,\text{given } a_l} &= E\{(f_{p,p,e}^i - f_{p,p,d}^i)^2\} = E\{(\sum a_l e_{n-1}^{u_l(i)})^2\} \\ &= \sum_l a_l^2 E\{(e_{n-1}^{u_l(i)})^2\} + \sum_{l,k,l\neq k} a_l a_k E\{e_{n-1}^{u_l(i)} e_{n-1}^{u_k(i)}\} \\ &= (\sum_l a_l^2 + \rho \sum_{l,k,l\neq k} a_l a_k) D_{c,n-1} \end{aligned}$$

In going from line 2 to line 3 in the above equation, we have assumed that the correlation coefficients between errors in every two neighboring pixels are the same, represented by ρ.

The fact that different values of $L_{p,p}, a_l$ are used for pixels in different P-MBs can be taken into account by taking the average of the factors $\sum_{l=1}^{L_{p,p}} a_l^2 + \rho \sum_{l,r,l\neq k} a_l a_k$ used over all P-MBs in the frame, and denote the average value by

$$a = E_a\{\sum_{l=1} a_l^2 + \rho \sum_{l,k,l\neq k} a_l a_k\} \tag{2}$$

Assuming this average value is the same in different frames, we have

$$D_{\mathrm{PR},n} = a D_{c,n-1}. \tag{3}$$

If an MB is lost, regardless of its coding mode, it will be concealed using temporal concealment with an estimated MV. Generally, the estimated MV may also be a non-integer vector, and the concealed value can be denoted by $f_{\mathrm{ECP}}^i = \sum_{l=1}^{L_{c,p}} h_l \tilde{f}_{n-1}^{s_l(i)}$, with $L_{c,p}$ and $s_l(i)$ differing from $L_{p,p}$ and $u_l(i)$ in general, and $\sum_l h_l = 1$. The average channel distortion, averaged over locations with different h_l, is

$$\begin{aligned} D_{\mathrm{L},n} &= E\{(\hat{f}_n^i - f_{\mathrm{ECP}}^i)^2\} = E\{(\hat{f}_n^i - \sum h_l \tilde{f}_{n-1}^{s_l(i)})^2\} \\ &= E\{(\hat{f}_n^i - \sum h_l \hat{f}_{n-1}^{s_l(i)} + \sum h_l e_{n-1}^{s_l(i)})^2\} \\ &= D_{\mathrm{ECP},n} + h D_{c,n-1}, \end{aligned} \tag{4}$$

$$\text{with } D_{\mathrm{ECP},n} = E\{(\hat{f}_n^i - \sum h_l \hat{f}_{n-1}^{s_l(i)})^2\}, \tag{5}$$

$$h = E_a\{\sum_{l=1} h_l^2 + \rho \sum_{l,k,l\neq k} h_l h_k\}. \tag{6}$$

Note that $\sum_l h_l \hat{f}_{n-1}^{s_l(i)}$ would be the concealed value using the same estimated MV in the absence of error propagation. Therefore, $D_{\mathrm{ECP},n}$ represents the average distortion associated with a particular temporal concealment algorithm, *in the absence of error propagation from previous frames.* For example, if we use the simple copy-from-previous-frame algorithm, then $L_{c,p} = 1$, $s_1(i) = i$, and $D_{\mathrm{ECP},n} = E_a\{(\hat{f}_n^i - \hat{f}_{n-1}^i)^2\}$ is the mean squared difference between two successively coded frames. When going from the second to the third line in (4), we have assumed that the concealment error at frame n is uncorrelated with the channel-induced error in frame n-1.

Substituting (3) and (4) and $D_{\mathrm{IR},n} = 0$ into (1) yields

$$D_{c,n} = P D_{\mathrm{ECP},n} + \alpha_n D_{c,n-1}, \tag{7}$$

$$\text{with } \alpha_n = a(1-\beta_n)(1-P) + hP. \tag{8}$$

The above recursion formula tells us that $D_{c,n}$ is the sum of the concealment distortion in this frame and the propagated error from the previous frame, with α_n being a factor controlling the decay of error propagation. When only integer motion vectors are used for motion compensation and temporal concealment, the constants $a = 1, h = 1$. In that case, $\alpha_n = 1 - \beta_n(1-P)$ is equal to the percentage of pixels at which error propagation will continue. With non-integer motion vectors, $0 < a < 1$ and $0 < h < 1$, making α_n smaller. Therefore, the spatial filtering incurred by fractional-pel motion-compensated prediction and concealment has the effect of attenuating the temporal error propagation.

The definition of a in (2) assumes the correlation of channel-induced error in neighboring pixels is a constant. We have found that this correlation in fact decreases when the random intra rate increases so that we can write a as $a + (1-\beta_n)b$. Therefore a more accurate model (but requiring one more parameter) is to replace (8) by

$$\alpha_n = (a + (1-\beta_n)b)(1-\beta_n)(1-P) + hP. \tag{9}$$

The distortion model in [3] has the same form as (7,8) but with $h = 1$, because it assumes frame-copy for concealment. The constant a was introduced to account for the so-called motion randomness. The derivation here shows clearly the relation of a with the interpolation coefficients used for motion compensation with non-integer motion vectors. The model in [3] also assumes the concealment distortion is proportional to the frame difference square, $D_{\mathrm{ECP},n} = eE\{(f_n^i - f_{n-1}^i)^2\}$, where e is a model parameter. This assumption is only valid for the frame-copy error-concealment method. We assume $D_{\mathrm{ECP},n}$ can be measured by the encoder, by running the same error concealment method on selected sample MBs as the decoder does.

Case II: With Intra-Prediction

With non-constrained intra-prediction, for received I-MBs, the distortion is no longer zero because an I-MB may be predicted (directly or indirectly) from neighboring pixels that are coded in the inter mode. To analyze this case, we assume that a pixel f_n^i is predicted by a weighted sum of several previously coded neighboring pixels in frame n, denoted by $f_{p,i,e}^i = \sum_{l=1}^{L_{p,i}} c_l \hat{f}_n^{q_l(i)}$.

If an I-MB is received, the intra-predicted value at the decoder is $f_{p,i,d}^i = \sum_{l=1}^{L_{p,i}} c_l \tilde{f}_n^{q_l(i)}$. Generally, the neighboring pixels used to predict a current pixel may come from either I-MBs or P-MBs. We will call these neighboring pixels I-neighbors and P-neighbors, respectively. These pixels are also received because they belong to the same slice as the current MB. The average distortion of P-neighbors is $D_{\mathrm{PR},n}$ by definition. Denoting the average distortion of I-neighbors by

$D^{\text{past}}_{\text{IR},n}$, the distortion of the current I-MB can be written as

$$D^{\text{current}}_{\text{IR},n} = E\{(f^i_{p,i,e} - f^i_{p,i,d})^2\} = E\left\{\left(\sum c_l e^{q_l(i)}_n\right)^2\right\}$$
$$= c_I D^{\text{past}}_{\text{IR},n} + c_P D_{\text{PR},n}, \text{ with}$$
$$c_{\text{I/P}} = E_a\{\sum_{l:\text{I/P}-\text{neighbors}} c_l^2 + \rho \sum_{l,k:\text{I/P}-\text{neighbors},l\neq k} c_l c_k\}.$$

In deriving the above result, we have assumed the channel distortion in I-neighbors and that in P-neighbors are uncorrelated. *The above recursion on the distortion of I-MBs in successive pixel locations show that intra-prediction causes spatial error propagation within the same frame.* We will assume that this distortion quickly converges after a few I-MBs so that $D^{\text{current}}_{\text{IR},n} = D^{\text{past}}_{\text{IR},n} = D_{\text{IR},n}$, and

$$D_{\text{IR},n} = \frac{c_P}{1-c_I} D_{\text{PR},n} = \frac{a c_P}{1-c_I} D_{c,n-1}.$$

Because c_P is likely to be proportional to $(1-\beta_n)$, we can assume that $\frac{ac_P}{1-c_I}$ is linearly proportional to $(1-\beta_n)$ and write

$$D_{\text{IR},n} = c(1-\beta_n) D_{c,n-1}. \tag{10}$$

The factor c can be larger or smaller than a, depending on the relative magnitude of c_I vs. c_P. This means that non-constrained intra-prediction can make the temporal error propagation in I-MBs worse than that in P-MBs.

Substituting (10),(3) and (4) into (1) yields the same recursion as in (7) but with

$$\alpha_n = (1-P)(1-\beta_n)(a+\beta_n c) + hP. \tag{11}$$

The proceeding analysis assumed unconstrained intra-prediction. With constrained intra-prediction as in H.264, only intra-coded neighboring pixels in the same slice can be used for intra-prediction so that $c_P = 0$. Consequently $c = 0$ and the overall distortion stays the same as in Case I.

Case III: With Deblocking Filtering

In the H.264 standard, deblocking filtering is applied in the so-called "in-place" manner, so that the filtered value for a pixel can be used for filtering following pixels. Let $\check{f}^i_n$ represent the reconstructed value for pixel f^i_n before filtering, and $\hat{f}^i_n$ the reconstructed value after filtering. Mathematically, we can describe the deblocking operation by $\hat{f}^i_n = \sum_{l:\text{past}} w_l \hat{f}^{r_l(i)}_n + \sum_{l:\text{future}} w_l \check{f}^{r_l(i)}_n$. The filter coefficients w_l are location and content dependent, satisfying $\sum_l w_l = 1$. In the decoder, if an MB is received, the same filtering is applied to the decoded values $\bar{f}^i_n$, with the filtered value $\tilde{f}^i_n = \sum_{l:\text{past}} w_l \tilde{f}^{r_l(i)}_n + \sum_{l:\text{future}} w_l \bar{f}^{r_l(i)}_n$. The average distortion for a received MB is

$$D^{\text{current}}_{R,n} = w_{past} D^{\text{past}}_{R,n} + w_{\text{future}} \bar{D}_{R,n}$$
$$w_{\text{past/future}} = E_a\{\sum_{l:past/fut.} w_l^2 + \rho \sum_{l,k:past/fut.,l\neq k} w_l w_k\}.$$

Assuming the distortion quickly converges so that $D^{\text{current}}_{R,n} = D^{\text{past}}_{R,n} = D_{R,n}$, we have

$$D_{R,n} = w\bar{D}_{R,n}, \text{with } w = \frac{w_{future}}{1-w_{past}}$$

where $\bar{D}_{R,n}$ is the distortion for a received MB if no deblocking filtering is applied. Substituting (3) and (10) for $\bar{D}_{R,n}$ for P- and I-MBs, respectively, yields

$$D_{\text{PR},n} = a' D_{c,n-1}, \qquad \text{with } a' = w_P a, \tag{12}$$
$$D_{\text{IR},n} = c'(1-\beta_n) D_{c,n-1}, \qquad \text{with } c' = w_I c, \tag{13}$$

where w_I and w_P are the "w" constants corresponding to I- and P-MBs, respectively. In general, their values differ because different deblocking filters are typically applied for I- and P-MBs. For a lost MB, deblocking is typically not applied after concealment. Therefore, its distortion stays as (4). Hence, the average distortion has the same form as (7,11) but with a and c replaced by a' and c', respectively. Depending on the relative magnitude of w_{past} and w_{future}, w can be either smaller or greater than 1. Therefore, recursive deblocking filtering can either attenuate or exacerbate error propagation.

Model Simplification

The previous analysis assumes that $D_{\text{ECP},n}$ and β_n varies from frame to frame and can be measured accurately. A simplified model results if we assume these values stay fairly constant. Let D_{ECP} and β denote the average concealment distortion and intra-rate, then recursion (7,11) becomes

$$D_{c,n} = PD_{\text{ECP}} + \alpha D_{c,n-1}, \tag{14}$$
$$\text{with } \alpha = (1-P)(1-\beta)(a+\beta c) + hP. \tag{15}$$

Assume the first frame is coded in the I-mode and has a channel distortion of $D_{c,0}$. Applying (14) recursively yields

$$D_{c,n} = PD_{\text{ECP}}(1+\alpha+\cdots+\alpha^{n-1}) + \alpha^n D_{c,0} \tag{16}$$
$$= PD_{\text{ECP}} \frac{1-\alpha^n}{1-\alpha} + \alpha^n D_{c,0} \tag{17}$$
$$\approx \frac{P}{1-\alpha} D_{\text{ECP}} = D_c \text{ for } n \text{ large and } |\alpha| < 1. \tag{18}$$

Note that α depends on P, so that the converged value D_c is NOT a simple linear function of P. In the special case of $a = h = 1$ and $c = 0$, the converged value $D_c = \frac{P}{\beta(1-P)} D_{\text{ECP}}$.

IV. Verification of the Model

The H.264 codec with different encoding options are employed to test the proposed models for different cases. To examine Case I (no intra-prediction), we used "constrained intra prediction" option instead, which should follow the same model as for Case I according to our analysis. We used one slice per frame so that a lost slice leads to a lost frame. We encoded the first 4 sec. of two QCIF sequences, "foreman" and "football", at 15 f/s. The first frame is coded as an I-frame, while the remaining 59 frames are coded as P-frames using forced intra rates of 3/99, 9/99, and 33/99. A constant QP=28 is used. The P-frame data are subjected

to random frame loss at rates of 1%, 5%, 10% and 15%. For a given target loss rate and a forced intra rate, 500 loss traces are generated. The channel distortion for each frame is determined by averaging the distortion resulting from all loss traces. The decoder conceals a lost frame by copying from the previous reconstructed frame. Hence the parameter $h = 1$. The concealment distortion $D_{\mathrm{ECP},n}$ is simply the mean squared difference between two encoded frames, which are directly measured. To obtain the model parameters a, b, c for each test sequence, we apply least square fitting to the recursion formula (7) with corresponding α_n defined for different cases, using the data obtained at different loss rates and intra rates.

Figure 1 shows the average channel-induced distortion over all P-frames vs. the packet loss rate for Case I. The β values indicated on the figure are the average intra rates over all P-frames corresponding to different forced intra rates. "Model1" refers to (7,8), whereas "Model2" refers to (7,9). We see that "Model2" fits the experimental data quite well over the large range of loss rates and intra rates examined (less accurate at high loss rates). "Model1", with one less parameter, is less accurate, but still provides a quite good approximation. "Model1" and "Model2" are obtained by using the actual β_n and $D_{\mathrm{ECP},n}$. "Model2-avg-DECP" is computed by using the average intra-rate and concealment distortion over all frames, which is almost as accurate as "Model2". Therefore, the model can estimate the average distortion accurately even if we only know the average intra rate and concealment distortion.

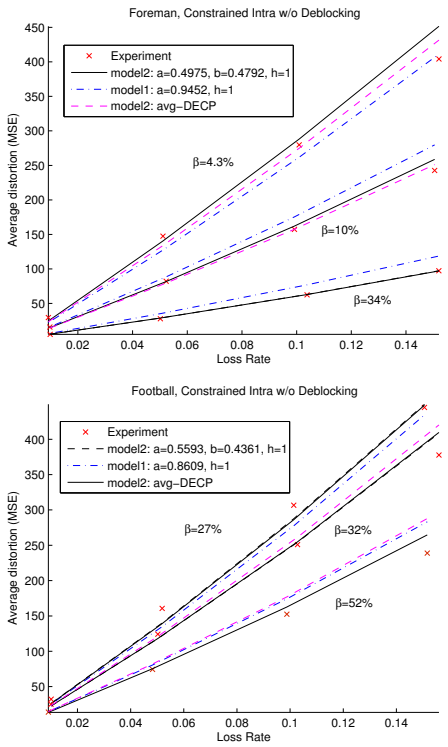

Fig. 1. Cases I: Constrained intra prediction without deblocking filtering.

Figure 2 shows the results for Case II. The model curve is computed using (7,11), and it fits the experimental data quite well, both with actual intra rate and concealment distortion, and their average values. For "foreman," the experimental curves corresponding to $\beta = 6.6\%$ and $\beta = 12\%$ are very close to each other, and the modelled curves fall on top of each other. Comparing Figs. 1 and 2, we see that under the same intra rate and packet loss rate, the channel distortion is much higher when non-constrained intra prediction is used. Therefore, for error resilience purpose, constrained intra prediction is much preferred.

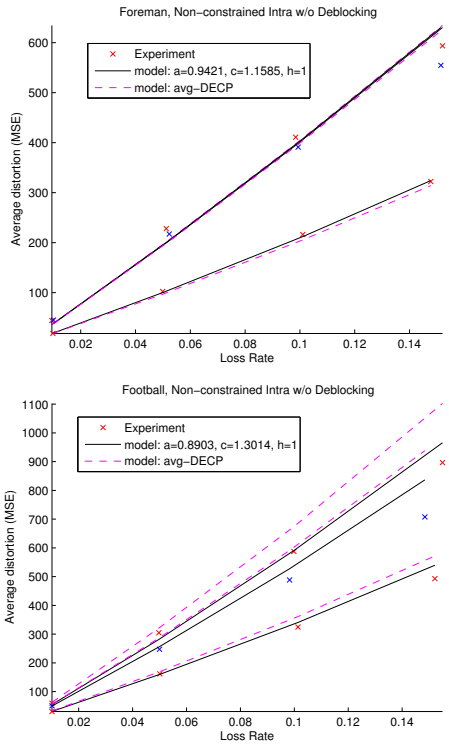

Fig. 2. Cases II: Unconstrained intra prediction without deblocking filtering.

With deblocking filtering, we obtained results very similar to Figs 1 and 2, when constrained and non-constrained intra-prediction are used, respectively, with slightly different model parameters.

The results shown here used the model parameters derived for the actual test sequences. We are conducting more simulations studies to see whether video sequences with similar motion characteristics have similar parameters. We are also validating the model when the decoder employs motion-compensated temporal concealment.

REFERENCES

[1] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," IEEE J. Select. Areas Commun., vol. 18, pp. 966-976, 2000.

[2] K. Stuhmuller, N. Farber, M. Link and B. Girod, "Analysis of video transmission over lossy channels", IEEE J. Select. Areas Commun., vol. 18, June 2000.

[3] Z. He, H. Cai, C. W. Chang, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding", CSVT, vol. 12, June 2002.