HYBRID END-TO-END DISTORTION ESTIMATION AND ITS APPLICATION IN ERROR RESILIENT VIDEO CODING

Xiaohui Wei[*]

Univ. of Texas at Arlington
Arlington, TX 76010, U.S.A.
xhwei@cse.uta.edu

Hua Yang and Jill M. Boyce

Corporate Research, Thomson Inc.
Princeton, NJ 08540, U.S.A.
{hua.yang2, jill.boyce}@thomson.net

ABSTRACT

This paper addresses the low complexity end-to-end distortion (ED) estimation problem for error resilient video coding. Unlike the existing "look-back-only" ED estimation paradigm, we propose a new hybrid paradigm involving both "look-back" and "look-ahead" estimation. For low complexity, our "look-back" estimation accurately accounts for the error propagation (EP) distortion from the last two frames only, while the impacts of "look-back" ignored frame losses are compensated by "look-ahead" frame-level EP approximation. The proposed hybrid scheme is then applied in ED-based RD optimization (ED-RDO) of both motion estimation (ME) and coding mode selection (MS). Results show that our hybrid estimation scheme yields accurate ED estimates, and when applied in ED-RDO ME and MS, significant performance gain is achieved over the other existing low complexity solutions.

Index Terms— end-to-end distortion, low complexity, error propagation, error resilience, video coding

1. INTRODUCTION

To achieve good video streaming performance, how to mitigate the packet loss impact from nowadays imperfect network transmission is a critical issue. At the encoder side, an efficient framework is end-to-end distortion (ED) based rate-distortion (RD) optimization, where how to accurately estimate ED is a challenging task.

Existing ED estimation schemes can be roughly categorized into pixel-based [1] [2], block-based [3] or frame-based [4] approaches. Accurate ED estimate may be achieved by the pixel-based ROPE method [1]. However, along with its high estimation accuracy, it also incurs a significant amount of computational complexity, which is not desirable in practical real-time video streaming systems. For low complexity, a simplified pixel-based distortion estimation (SPDE) approach was proposed in [2], where only two most likely loss events, (i.e. the loss of the last two frames respectively), are considered. However, ignoring all the other loss events greatly compromises the estimation performance. Alternatively, block-based schemes, e.g. [3], also reduce the complexity of pixel-based ROPE estimation roughly by a factor of the block-size (e.g. 16 as for 4x4 blocks). However, the estimation accuracy is also greatly compromised due to the reduced resolution. Frame-level ED estimation [4] targets ED estimation of a whole frame. In this case, all the complicating factors such as Intra coded macro-blocks (MB), sub-pixel prediction, deblocking filtering, etc. may be respectively modelled with one single parameter for each frame, and thus, the whole estimation involves neglectable computation complexity. In practice, frame-level estimation is usually applied in frame-level ED based RD optimization (ED-RDO) problems, while for the concerned MB-level ED-RDO tasks, such as ED-RDO ME and MS, either pixel- or block-based ED estimation is required.

In this work, we intend to seek a low complexity ED estimation solution, which renders a better trade-off between estimation accuracy and complexity, and is applicable in ED-RDO ME and MS. Herein, we assume the group-of-picture (GoP) based video coding framework, which is commonly applied in practical video streaming systems. Specifically, motivated from the GoP-level ED estimation of the existing FODE method [5], we propose a novel scheme called HEED (hybrid estimation of ED). Similar as in SPDE [2], HEED also accounts for error propagation (EP) distortions only for the respective loss of the last two frames. Differently, instead of excluding the distortion contributions from all the other loss events, we take their impacts into account, and use frame-level approximation to estimate the EP distortions from the current frame to all the remaining frames of the GoP. Moreover, we emphasize that while SPDE can be regarded as a simplification of ROPE, the proposed HEED scheme is, however, simplified from FODE, as will be more clearly explained in the following sections. As such, HEED takes both the high accuracy benefit from pixel-based calculation of past EP distortion, and the low complexity benefit from frame-level approximation on future EP distortion. Simulation results show that in spite of the simplification, HEED yields fairly accurate GoP-level distortion estimate. When applied in ED-RDO ME

[*] This work is done when Xiaohui Wei was an Intern at Thomson Inc.

1-4244-0728-1/07/$20.00 ©2007 IEEE

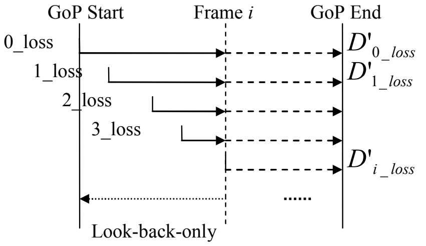

Fig. 1. The existing "look-back-only" ED estimation paradigm.

and MS, it significantly outperforms the other existing low complexity solutions.

2. HYBRID ESTIMATION OF END-TO-END DISTORTION

Let us first take a look at ED of the whole GOP, denoted by $E\{D_{GoP}\}$. In [5], a scheme called FODE (first order distortion estimate) was proposed, which approximates $E\{D_{GoP}\}$ with its first order Taylor expansion. In practice, the packet loss rate addressed by error resilient video coding is not large, e.g. $p < 10\%$. Beyond that, one has to use FEC or other techniques to effectively reduce p itself. With small p, the FODE model is fairly accurate. Its MSE $E\{D_{GoP}\}$ estimate is as follows.

$$E\{D_{GoP}\}\simeq D_{no\_loss} + p \cdot \sum_{i=0}^{N-1} \gamma_i. \tag{1}$$

Herein, N is the GoP size, and D_{no_loss} denotes the GoP distortion without any packet loss, i.e. the source coding distortion only. Throughout the paper, for simplicity, we assume data of one frame is packetized into one packet. γ_i is the 1st order Taylor expansion coefficient of frame i, which can be expressed as

$$\begin{aligned}
\gamma_i &= D_{i\_loss} - D_{no\_loss} && (2)\\
&= \sum_{j\geq i}\sum_{k=0}^{A-1}[(\tilde{f}^k_{j,i\_loss} - f^k_j)^2 - (\hat{f}^k_j - f^k_j)^2] && (3)\\
&\simeq \sum_{j\geq i}\sum_{k}(\tilde{f}^k_{j,i\_loss} - \hat{f}^k_j)^2 && (4)\\
&= D'_{i\_loss}. && (5)
\end{aligned}$$

Herein, A is the frame size. f^k_j and $\hat{f}^k_j$ represent the original and encoder reconstructed (i.e. no loss case) values of pixel k in frame j. D_{i_loss} and $\tilde{f}^k_{j,i\_loss}$ denote the GoP distortion and the decoder reconstructed pixel values, when *only* frame i is lost. In (4), the approximation is due to the omission of correlation terms. D'_{i_loss} denotes the EC and EP distortions due to the loss of frame i. (The prime is to indicate that the reference here is $\hat{f}^k_j$, but not f^k_j.) From the above equations, we can see that $E\{D_{GoP}\}$ can be regarded as a linear combination of D'_{i_loss}, as is illustrated in Fig. 1.

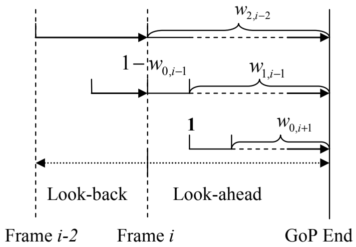

Fig. 2. The proposed HEED approach.

Although FODE was originally proposed to address the optimization problems of *coded* video, we emphasize that its simple linear representation of $E\{D_{GoP}\}$ renders useful insight as well on the concerned MB-level optimization tasks in the *encoding* process. In this case, when it comes to encoding a particular frame, one needs to identify for each MB their respective importance in terms of how they affect $E\{D_{GoP}\}$. For this, most, if not all, of the existing ED estimation approaches are "look-back-only" methods, where for the current frame, besides its own EC distortion, they basically estimate the overall *past* EP distortion due to the respective loss of each one of the previous frames in the GoP, as illustrated in Fig. 1. While the optimal ROPE approach accurately accounts for EP distortion from *all the past* individual frame loss events, to reduce complexity, the SPDE approach only considers EP distortion from the loss of each one of the *last two* frames. We emphasize that when applied in optimizing coding decisions of frames, this "look-back-only" paradigm still yields good inter-frame synergy, on condition that each frame can accurately estimate past EP distortion, as is the case for the optimal ROPE estimate. Because, in that case, when optimizing the coding decisions of the current frame, there is no need to worry about incurred future EP distortion, as that will be accurately considered in the following frames' optimizations. However, in the SPDE case, each frame only considers limited EP effect from the last two frames, which means that beyond the following two frames, EP distortion from the current frame will be completely ignored in optimizations of the remaining frames. In this case, the "look-back-only" paradigm cannot render good inter-frame synergy any more.

In light of the above analysis and motivated from the FODE distortion model, we propose a novel hybrid low complexity ED estimation approach, i.e. HEED. Similar as in SPDE, HEED also considers for each pixel the exact past EP distortion up to the 2nd last frame. However, instead of completely ignoring the impacts of all the other frame loss events, in HEED, we use frame-level EP factor approximation to explicitly account for the EP distortion from the current frame to all the remaining frames in the GoP, and thus, yield a hybrid paradigm involving both pixel-level "look-back" and frame-level "look-ahead" estimation, as illustrated in Fig. 2. Herein, $w_{0,i}$ denotes the weighting factor for considering the frame i loss EP branch *right at the loss of frame* i, while $w_{1,i}$ and $w_{2,i}$ denote the weighting factors for considering the frame

i loss EP branch respectively *at one frame or two frames after frame i loss*. Note that the three weighting factors of the same i should be summed up to 1, such that: the *complete* EP branch of the loss of each particular frame is *equivalently counted exactly once* in ED estimation of all the frames in the GoP. In that case, summing up the estimated ED over all the frames will yield an accurate estimate for ED of the whole GoP, as shown later in our simulation results.

Next, we describe how to conduct HEED ED estimation at each particular frame. In this work, we assume motion-copy error concealment at the decoder, where when a frame is lost, motion-vectors (MV) from collocated MBs in the previous frame is used to conceal the current frame via motion compensation. (Details of motion-copy EC refer to [6].) As such, the MV or coding mode of the current frame MB will also affect the EC distortion of the collocated MB in the next frame. Assuming the MB containing pixel k in frame i is *Inter* coded, our HEED method estimates ED of the pixel as

$$E\{D_i^k\} = D_{i,no\_loss}^k + p{\cdot}D_{EP,i}^k, \tag{6}$$

where

$$D_{EP,i}^k = D_{EP,i,i-2\_loss}^k + D_{EP,i,i-1\_loss}^k + D_{EP,i+1,i+1\_loss}^k. \tag{7}$$

The three right-hand side items of (7) correspond to the three considered EP branches in Fig. 2, respectively, which can be expressed as follows.

$$D_{EP,i,i-2\_loss}^k = w_{2,i-2}{D'}_{i,i-2\_loss}^k(1+\alpha_{1\to(N-i-1)}), \tag{8}$$

$$D_{EP,i,i-1\_loss}^k = {D'}_{i,i-1\_loss}^k(1-w_{0,i-1}+w_{1,i-1}\alpha_{1\to(N-i-1)}), \tag{9}$$

$$D_{EP,i+1,i+1\_loss}^k = {D'}_{i+1,i+1\_loss}^k(1+w_{0,i+1}\alpha_{1\to(N-i-2)}), \tag{10}$$

where,

$$\alpha_{1\to N} = \alpha + \alpha^2 + ... + \alpha^N. \tag{11}$$

Herein, α denotes the EP factor of a frame. We emphasize that modelling the ED effect of a frame with one single factor is a commonly adopted practice in existing frame-level ED estimation schemes for low complexity, where the overall factor α may involve various factors that accounts for Intra MBs, sub-pixel prediction, Intra-prediction, and de-blocking filtering, respectively, as discussed in [4]. In this work, for simplicity, $\alpha = 1 - \beta$, where β denotes the Intra MB percentage of a frame. On the other hand, the EP distortions from the last two frames are exactly calculated, whose resultant accuracy is even higher than that of the optimal ROPE approach, as we go through exactly the same EC and reconstruction process as the decoder would do when a frame is lost. However, similar exact calculation is impossible for the next frame EC distortion, as $\hat{f}_{i+1}^k$, and sometimes even $\hat{f}_i^k$, are not available at the time of coding frame i. In this work, we approximately estimate this term using the original references of frame i and $i+1$. Also, note that if a pixel is in a *Intra* coded MB, there will be no EP distortion terms from the last two frames, and only next frame EC distortion term stays in the above equations. Herein, we omit these equations.

In HEED, a critical issue is how to determine the involved weighting factors w_0, w_1, and w_2. Firstly, using one single EP factor α to model the actual complicated EP process is not accurate. Hence, it is desirable to evenly distribute the weight among the three factors. In that case, the overall modelling error will be reduced via averaging over the three items. Secondly, our HEED estimation will be applied in the concerned ED-RDO ME and MS problem. With the assumed motion-copy EC at the decoder, when a previous frame collocated MB is an Intra-MB, it will be treated the same as a Skip-MB, and the median MV from neighboring MVs will be used for concealment. In this case, although Intra coding of the current frame MB effectively stops existing EP from the past, it may as well incur more next frame EC distortion and hence more resultant EP distortion in the following frames than Inter mode coding, as an Inter mode has more flexibility to find a better MV so as to yield lower next frame EC distortion. Hence, the ratio between w_0 and $w_1 + w_2$ will directly affect the important Intra/Inter mode selection, and hence the overall ED-RDO performance. From experiments, a desirable strategy is to give more weight to w_0 for the beginning frames in a GoP and less weight to it for the ending frames. Finally, our adopted weighting factor setting is

$$w_{0,i} = \frac{N-i-1}{N},\ w_{1,i} = w_{2,i} = \frac{1}{2}(1-w_{0,i}). \tag{12}$$

The resultant $E\{D_i^k\}$ is then applied in both ED-RDO ME and MS, where the optimization problem is commonly formulated as minimization of a certain Lagrangian cost. Herein, we omit the details. One comment on our ED-RDO ME scheme is that instead of the common sum of absolute difference, MSE of the prediction residue is used to replace the D_{i,no_loss}^k in (6).

3. SIMULATION RESULTS

Our simulation is based on a proprietary H.264/AVC Baseline Profile encoder of Thomson Inc., where the Lagrangian minimization framework for RDO ME and MS is the same as that in the JM reference encoder. All the sequences are $30f/s$ and coded into GoPs with each GoP containing 30 frames. In experiment, only constrained Intra-prediction and single reference frame is enabled. All the various MB coding modes, sub-pixel prediction, de-blocking filtering, etc. of H.264/AVC are enabled. For simplicity, we assume no packet loss for the I-frames. For each packet loss rate p, 300 randomly generated packet loss patterns were applied at the decoder, and the average distortion or PSNR is computed. The encoder assumed the same exact value of p in its HEED calculation.

We first evaluate the GoP-level ED estimation performance of HEED. Herein, sequences are coded with fixed $QP = 33$.

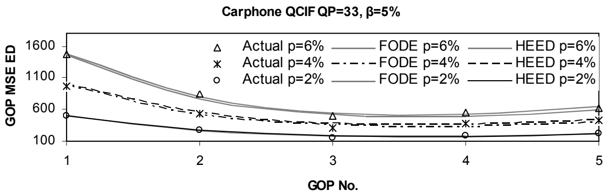

Fig. 3. GoP ED estimation performance.

Table 1. Performance with various sequences and bit rates. Except CIF Foreman, all the others are QCIF sequences.

$p = 3\%$	kb/s	Conv.	Forced I	SPDE	HEED
News	128	35.12	34.76	36.55	36.85
	256	36.49	36.91	40.11	40.44
Carphone	128	33.21	33.46	33.96	34.89
	256	34.14	34.80	35.83	37.07
Stefan	256	27.69	27.86	27.91	27.99
	384	28.82	29.10	29.46	29.87
Foreman	384	32.86	32.91	33.60	33.82
	512	33.41	33.58	34.38	34.70

In the coding of each P-frame, a fixed percentage of MBs are randomly selected to be Intra coded (denoted by β), and the exact same value of β is used to calculate α in HEED estimation. In this case, The result of the QCIF "Carphone" sequence with $\beta = 5\%$ is shown in Fig. 3. Herein, "Actual" denotes the averaged decoder distortion to represent the actual value of ED estimate. We can see that HEED achieves fairly accurate GoP ED estimate at all the testing packet loss rates, and its performance is quite similar to the existing FODE approach. Similar results are also obtained for many other sequences, and the observed relative estimation error is usually below 3%. Note that HEED can be regarded as a simplified version of FODE. In spite of that, the simplification does not yield much estimation accuracy drop.

We then evaluate the performance of applying HEED in ED-RDO ME and MS. For comparison, we tested the other two existing low complexity solutions, including: the SPDE approach (denoted by "SPDE"), and the naive forced Intra MB coding approach (denoted by "Forced Intra"), where for packet loss rate p, $100{\cdot}p$ percent MBs in each frame are randomly selected for Intra coding. We also include the coding results with conventional RDO ME and MS (denoted by "Conventional"). Note that unlike most of the existing work, our coded sequences involve periodic I-frames due to the GoP structure, which already yield a certain degree of error resilience.

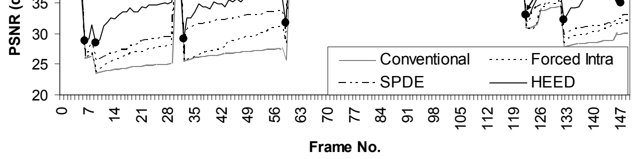

Fig. 4. Performance with one specific packet loss realization.

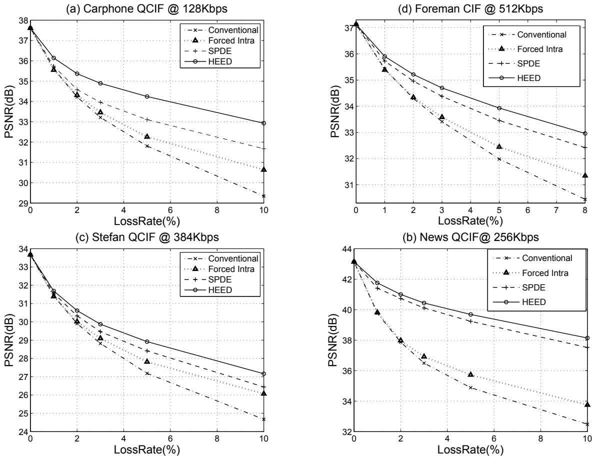

Fig. 5. Performance with various packet loss rates.

The extensive testing results of various sequences, bit rates, and various packet loss rates are thoroughly summarized in Table 1 and Fig. 5, respectively. We can see that in all the tested cases, the proposed ED-RDO ME and MS scheme with HEED significantly outperforms all the other methods. Specifically, the PSNR gain of HEED over SPDE reaches $1.29dB$ and is $0.47dB$ on average. Fig. 4 shows the PSNR vs. frame number curves of the "Carphone" sequence. We can see that comparing with all the other methods, "HEED" always yields not only lower EC distortions at the time of frame loss, but also lower EP distortions afterwards. All these results substantially proves the effectiveness of the proposed HEED-based error resilient video coding solution.

4. REFERENCES

[1] R. Zhang, S. L. Regunathan and K. Rose, "Video coding with optimal intra/inter-mode switching for packet loss resilience". *IEEE Journal Select. Areas Commun.*, vol. 18, no. 6, pp. 966-76, 2000.

[2] T. Wiegand, N. Farber, K. Stuhlmuller, and B. Girod, "Error-resilient video transmission using long-term memory motion-compensated prediction," *IEEE Journal Select. Areas Commun.*, vol. 18, no. 6, pp. 1050-62, June 2000.

[3] G. Cote and F. Kossentini, "Optimal intra coding of blocks for robust video communication over the Internet," *Sig. Processing: Image Commun.*, pp. 25-34, vol. 15, Sept. 1999.

[4] Y. Wang, Z. Wu and J. Boyce, "Modeling of Transmission-Loss-Induced Distortion in Decoded Video," *IEEE Trans. on Circuits Syst. Video Tech.*, vol. 16, no. 6, pp. 716-32, Jun. 2006.

[5] R. Zhang, S. L. Regunathan, K. Rose, "End-to-end distortion estimation for RD-based robust delivery of pre-compressed video," *35th Asilomar Conf.*, vol. 1, pp. 210-14, 2001.

[6] M. C. Hong, L. Kondi, H. Scwab, and A. K. Katsaggelos, "Error Concealment Algorithms for Compressed Video," *Sig. Processing: Image Commu.*, vol. 14, pp. 437-92, 1999.