COMPLEXITY SCALABLE HYBRID END-TO-END DISTORTION ESTIMATION FOR CONVERSATIONAL VIDEO STREAMING

Hua Yang, Xiaohui Wei, and Jill M. Boyce*

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

ABSTRACT

In conversational video streaming applications, periodic I-frame coding is not allowed due to strict low delay requirement. For this scenario, we propose a novel hybrid paradigm, which takes advantage of both accurate pixel-level estimation and efficient block-level estimation. Specifically, error propagated (EP) distortion from the last N coded frames are exactly calculated per pixel for high accuracy, while EP effect beyond the frame limit N are estimated per macroblock for low complexity. By varying the frame limit, flexible tradeoffs between complexity and performance can be achieved, rendering our scheme especially suitable for wireless video streaming, where stringent power resource is always a challenge. Simulation results justify the effectiveness of the proposed hybrid ED estimation scheme.

Index Terms— end-to-end distortion, complexity scalability, error resilience, conversational video streaming

1. INTRODUCTION

End-to-end distortion (ED) based rate-distortion (RD) optimization (ED-RDO) is an efficient and general framework to improve error resilience of video streaming. In live video streaming applications, ED-RDO can be applied to optimize various encoding parameters/options, such as macroblock (MB) coding modes, quantization step sizes, prediction references, or motion vectors (MV), etc. In these techniques, how to accurately estimate the ED is a critical and challenging task. Existing solutions suggest either pixel-based [1] or block-based [2] approaches. Accurate ED estimate may be achieved by the pixel-based ROPE method [1]. However, it incurs significant increase on complexity. To reduce complexity, a simplified pixel-based distortion estimation (SPDE) approach was proposed in [3], where only two most likely loss events, (i.e. the respective loss of the last two frames), are considered. However, ignoring all the other loss events greatly compromises the estimation performance. Alternatively, block-based schemes, e.g. [2], reduce complexity via reducing the resolution of ED estimation. However, the estimation accuracy is also compromised.

In this work, we propose a novel hybrid ED estimation paradigm, which takes advantage of both accurate pixel-level estimation and efficient block-level estimation. Specifically, error propagated (EP) distortion from the last N coded frames are exactly calculated per pixel for high accuracy, while EP effect beyond the frame limit N are estimated per MB for low complexity. By varying the frame limit N, different tradeoffs between complexity and performance can be achieved. This complexity scalability renders more flexibility on system design, which is especially useful in wireless/mobile video streaming applications, where stringent power resource (due to limited battery life) is always a practical limitation on the overall system performance.

We identify that both the proposed hybrid scheme and the other existing pixel- or block-base schemes are all "look-back-only" approaches, meaning that the current frame ED is estimated simply by accounting for error propagation from all the past frames. We emphasize that this "look-back-only" paradigm is effective mostly in *conversational* streaming applications, e.g. video conferencing or telephony. These two-way communication applications strictly require low delay, and hence, there is no periodic I-frames or the commonly known group-of-picture (GOP) structure applied. (Because the transmission time of an I-frame is usually much longer than that of a P-frame.) On the other hand, in *non-conversational* streaming, e.g. video-on-demand or broadcasting, delay requirement is relaxed due to one-way communication. Moreover, these applications usually require fast random access to coded video to enable, e.g. fast audience joining, fast channel change, fast forward/backward, etc. Therefore, periodic I-frame coding or the GOP structure is commonly applied. In this case, when estimating the current frame ED for encoding ED-RDO, only considering ED from the past frames may not be an effective strategy, because it ignores the fact that there will be a coming I-frame after the end of the current GOP, which already provides a certain level of error resilience. In

*Xiaohui Wei is with Univ. of Texas at Arlington, Arlington, TX 76010, U.S.A.

1-4244-1437-7/07/$20.00 ©2007 IEEE

fact, as for GOP-based video coding, it is more effective to consider not only past frame EP but also *future EP effect before the GOP end*. For this, we already proposed in [4] another type of hybrid scheme, involving both "look-back" and "look-ahead" estimation. In contrast, the proposed scheme herein is different, and particularly targeting non-GOP-based video coding.

2. HYBRID DISTORTION ESTIMATION FOR NON-GOP-BASED VIDEO CODING

2.1. Existing First Order Distortion Estimate (FODE)

Our proposed scheme is originated from an existing approach called FODE (first order distortion estimate) in [5], which was originally proposed to approximate GOP-level ED estimation for pre-compressed video streaming. Let $E\{D_{GOP}\}$ denote ED of a GOP, which is approximated in FODE by its first order Taylor expansion. In practice, the packet loss rate (denoted by p) 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 has proved to be fairly accurate [5]. Its mean squared error (MSE) $E\{D_{GOP}\}$ estimate is as follows.

$$E\{D_{GOP}\} \simeq D_{no\_loss} + p \cdot \sum_{i=0}^{M-1} \gamma_i. \qquad (1)$$

Herein, M is the GOP size, and D_{no_loss} denotes the GOP distortion without any packet loss, i.e. the source coding distortion. 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 indicates that the reference here is $\hat{f}^k_j$, but not f^k_j.)

From the above equations, we can see that a nice property of FODE is that $E\{D_{GOP}\}$ is expressed as a *linear* combination of D'_{i_loss}, as is also illustrated in Fig. 1. In the case of multiple frame losses, FODE approximation is actually equivalent to assuming that their respective EP effects are *linearly additive*.

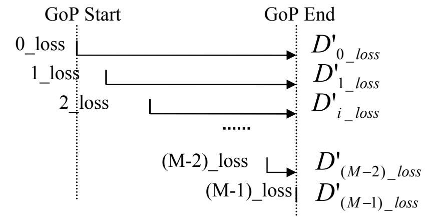

Fig. 1. The existing FODE estimation paradigm.

2.2. The Proposed Hybrid Distortion Estimate

Although FODE was originally proposed to address the ED estimation problem in *GOP-based video coding*, we found that its linear additive approximation is indeed general, and can be extended to our concerned non-GOP-based video coding scenario, which is equivalent to the case of *infinite GOP size*. In our problem, we need to estimate ED for each frame, instead of for each GOP as in [5]. Knowing that periodic I-frames are not involved, at each frame, we only need to account for EP effects from all the past frames. Hence, ED of frame i can be approximated as:

$$E\{D_i\} \simeq D_{no\_loss,i} + p \cdot \sum_{j\leq i} \gamma_{j\rightarrow i}. \qquad (6)$$

Herein, $D_{no\_loss,i}$ is source coding distortion of frame i. $\gamma_{j\rightarrow i}$ denotes EP distortion from the lost frame j to the current frame i, which is defined as:

$$\gamma_{j\rightarrow i} = \sum_{k}(\tilde{f}^k_{i,j\_loss} - \hat{f}^k_i)^2. \qquad (7)$$

Note that, as shown in (4), original FODE suggests pixel-level calculation for all the involved $\gamma_{j\rightarrow i}$ terms, which, however, entails calculating and storing a pixel-level EP distortion map for each single frame loss event. Consequently, both the computational complexity and storage cost will linearly increase with the number of coded frames increased. Therefore, in order to render (6) feasible, we propose a hybrid estimation approach, where exact pixel-level EP distortion is only calculated up to a certain limited number (denoted by N) of most recently coded frames, while for the past frames beyond the limit, i.e. $j > N$, we only maintain *one single MB-level distortion map* to jointly capture their overall EP impact. The MB-level calculation is defined as follows.

$$\sum_{j<N} \gamma_{j\rightarrow i} \simeq D_{out\rightarrow i} = \sum_{mb} D_{out\rightarrow i,mb}. \qquad (8)$$

Herein, $D_{out\rightarrow i}$ denotes the EP distortion from all the beyond-limit past frames to the current frame i, where the contribution of any individual MB (indexed by mb) is identified by

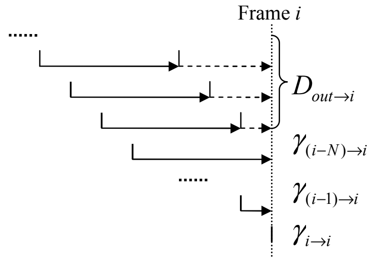

Fig. 2. The proposed hybrid estimation paradigm for non-GOP-based video coding.

$D_{out\to i,mb}$, and calculated as:

$$D_{out\to i,mb} = \beta(mv) \sum_{k\in\Omega(mv)} \alpha_k \cdot D_{map,i-1,k}. \qquad (9)$$

Herein, for clarity, we only assume one MV per MB. However, in practice, the same calculation is indeed applicable for all the eligible sub-MB/block options of H.264/AVC. mv denotes its MV. $\Omega(mv)$ is the set of indices of the MBs in frame $i-1$ that are overlapped with the prediction reference area of the current MB. α_k denotes the ratio of the number of overlapped pixels with MB k of frame $i-1$ against the total number of pixels in a MB. $\beta(mv)$ is a factor capturing spatial filtering effects from 1/2-pel or 1/4-pel prediction. In practice, β is set to 1 for full-pel, and heuristically, 0.98 and 0.96 for 1/2- and 1/4-pel, respectively. $\{D_{map,i-1,k}\}$ represents the MB-level distortion map at frame $i-1$, which is updated from frame to frame as follows.

$$D_{map,i,mb} = D_{out\to i,mb} + \gamma_{(i-N)\to i,mb}. \qquad (10)$$

The proposed scheme is also illustrated in Fig. 2. Herein (and also in Fig. 1), we use solid lines to indicate pixel-level calculation, while dashed lines for MB-level calculation.

Clearly, with MB-level calculation of $D_{out\to i}$ defined in above, FODE can be extended to the concerned infinite GOP size case with feasible complexity. Furthermore, by varying the frame limit N, different levels of complexity vs. performance trade-off may be achieved. The resultant complexity scalability allows for more flexibility, and is highly desirable in wireless/mobile video streaming applications, where the limited battery life, and hence power resource, will inevitably affect the overall system performance. With our scheme, one can adaptively change N, and hence, the incurred complexity, according to differing power capabilities with differing devices or in differing channel conditions, or according to differing video characteristics, etc. In this way, a better overall resource cost and system performance trade-off may be achieved.

Comparing with SPDE, which completely ignores the impact of all the beyond-limit frames, our hybrid scheme uses MB-level estimation for that, which greatly improves the accuracy, as will be shown in Section 3. Comparing with existing pixel-based ROPE [1] or block-based [2] schemes, we notice that all these schemes conduct accurate *recursive* frame-by-frame calculation of distortion maps. Without FODE approximation, these schemes may respectively yield better estimation performance than the corresponding pixel-level or MB-level estimations in our scheme. However, this also prevents them from rendering complexity scalability.

2.3. ED-RDO With Hybrid Distortion Estimate

The proposed hybrid ED estimate is generally applicable in all the existing ED-RDO video coding techniques. As a particular example, in this work, we apply it in ED-RDO motion estimation (ME) and coding mode selection (MS). The optimization problem is commonly formulated as: independently selecting the best MV and coding mode for each MB/block to minimize a Lagrangian cost that weighs ED versus bit rate by a certain Lagrangian multiplier. 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 keep accordance with the estimated MSE ED. Another comment is that, throughout this work, we assume motion-copy error concealment (EC) at the decoder, where when a frame is lost, MVs 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.

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 P-frames except for the 1st I-frame. In experiment, only constrained Intra-prediction and single reference frame is enabled, and de-blocking filtering is excluded. All the various MB coding modes and sub-pixel prediction of H.264/AVC are enabled. For simplicity, we assume no packet loss for the 1st I-frame. 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 ED calculation. Herein, "Actual", "FODE", and "SPDE_2" denote the decoder actual ED, FODE estimated ED and SPDE estimated ED with the frame limit 2, respectively. "Scalable_0"–"Scalable_20" represent our proposed complexity scalable scheme with different frame limits.

We first evaluate the frame-level ED estimation performance. In each P-frame, a fixed percentage of MBs are randomly selected to be Intra coded (denoted by "Intra ratio"). The results are shown in Fig. 3. In Fig. 3 (a), estimation performance is measured in relative frame ED estimation error

Table 1. ED-RDO PSNR performance with various sequences. $p = 5\%$, Stefan: $256kb/s$, others: $128kb/s$.

	Conventional	Forced Intra	SPDE_2	Scalable_0	Scalable_1	Scalable_2	Scalable_5	Scalable_10
Miss_America	39.35	41.07	40.03	41.14	41.16	41.13	41.12	41.32
Carphone	28.60	32.13	31.19	33.54	33.64	33.75	33.83	33.95
News	33.12	33.58	35.49	36.17	36.21	36.31	36.33	36.50
Foreman	29.86	31.42	30.71	31.69	31.83	31.85	31.99	32.15
Stefan	24.74	27.26	25.72	27.83	28.09	28.23	28.31	28.35

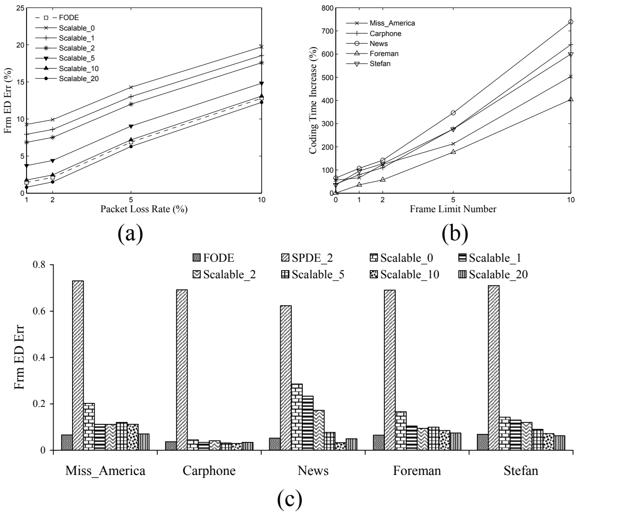

Fig. 3. Frame-level ED estimation performance. Intra ratio= 10%, Relative frame ED error performance: (a) at various loss rates, Stefan, $256kb/s$; (c) for various sequences, $p = 5\%$, Intra ratio= 10%, Stefan: $256kb/s$, others: $128kb/s$. (b) Relative coding time increase: the same conditions as in (c).

(denoted by "Frm ED Err"), defined by $|D_{est,i} - D_{act,i}|/D_{act,i}$, where $D_{est,i}$ and $D_{act,i}$ respectively denote encoder estimated and decoder actual ED of frame i. We can see that, in general, with increased frame limit, our scheme achieves more accurate ED estimate, and the performance more closely approaches that of FODE. Similar observations can be also made on many other sequences as shown in Fig. 3 (c). Moreover, it is obvious that completely ignoring the EP impact from beyond the past two frames as SPDE_2 does seriously degrades the estimation accuracy. On the other hand, Fig. 3 (b) gives the increased total encoding time percentage with respect to conventional encoding (involving no ED estimation). One can see that with larger frame limits applied, along with increase ED estimation accuracy, the coding time is also increased.

Next, we evaluate the performance when applying our hybrid scheme in ED-RDO ME and MS. We also tested a naive approach with forced Intra ratio equal to the loss rate (denoted by "Forced Intra"). The conventional RDO ME and MS method is denoted by "Conventional". The results are summarized in Fig. 4 and Table 1. Herein, it is clear that the proposed Scalable_0–Scalable_20 always outperform the other approaches, and higher frame limit numbers generally gives higher PSNR performance.

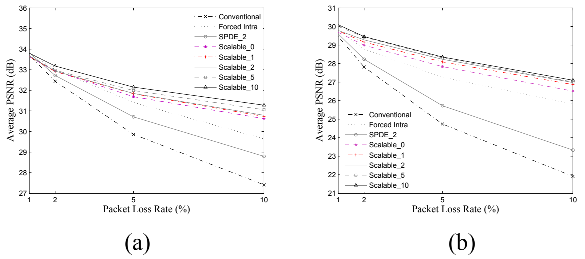

Fig. 4. ED-RDO performance at various packet loss rates. (a) Foreman, $128kb/s$. (b) Stefan, $256kb/s$.

In summary, all the above results substantiate the effectiveness of the proposed hybrid ED estimation approach. In practice, one can measure and compare the ED-RDO performance and the coding time increase curves under real system operating conditions, and based on that, select the most effective frame limit number.

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] 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.

[3] 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.

[4] X. Wei, H. Yang, and J. M. Boyce, "Hybrid end-to-end distortion estimation and its application in error resilient video coding," to appear in *ICASSP 2007*.

[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.