RECURSIVE END-TO-END DISTORTION ESTIMATION WITH MODEL-BASED CROSS-CORRELATION APPROXIMATION

Hua Yang and Kenneth Rose

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106

ABSTRACT

Accurate end-to-end distortion estimation is critical to efficient rate-distortion (R-D) optimization of encoder decisions for video transmission over lossy packet networks. This work focuses on extensions of the recursive optimal per-pixel estimate (ROPE), which has been shown to provide accurate end-to-end distortion estimation. Of particular interest are difficulties due to sub-pixel prediction and other pixel averaging operations, for which the existing ROPE encounters cross-correlation terms, whose exact estimation requires prohibitive storage and computational complexity. In this paper, we propose two model-based methods, which approximate the cross-correlation of two pixels as a function of their available first and second marginal moments. This allows an approximate extension of ROPE to handle sub-pixel prediction and other pixel averaging operations, at no additional storage cost, and no significant additional complexity. Simulations provide evidence for the performance gains of the proposed methods, and in particular, demonstrate that the resulting accuracy is very close to that of ROPE when it is optimal, i.e., in the case of full pixel prediction.

1. INTRODUCTION

A critical concern in video networking is how to adequately mitigate the impact of packet loss. Rate-distortion (R-D) optimization has been widely recognized as an efficient framework for incorporating error robustness tools in video coding system, and has been adopted in a variety of error resilient video coding techniques (see, e.g., [1][2]). While the coding bit rate is easily controlled by the encoder, the overall end-to-end distortion is much more elusive and must be accurately estimated. Much research effort has been dedicated to this problem in recent years, including [3][4]. Among existing schemes, the recursive optimal per-pixel estimate (ROPE) [4] demonstrates superior performance in that it accurately estimates the end-to-end distortion of decoder reconstruction by taking into account all the effects of quantization, packet loss, and error concealment. It has been frequently applied to R-D optimized mode selection in several video coding frameworks [4][5][6].

Most existing ROPE related approaches assume integer pixel motion compensation rather than sub-pixel prediction, which is known to provide superior compression performance and has long been adopted by video coding standards, such as MPEG-4, H.263 and H.26L [7]. The reason is that sub-pixel prediction involves bilinear interpolation, which gives rise to *inter-pixel cross-correlation* terms in the distortion estimation process of ROPE. Specifically, in order for ROPE to exactly calculate its estimate, it is necessary to compute and store all inter-pixel cross-correlation values in the frame. This entails prohibitive amount of computation and storage space, and hence makes it impractical to implement ROPE in a video coding system with sub-pixel prediction. Moreover, *pixel averaging operations*, which causes those cross-correlation terms, appears not only in sub-pixel prediction but also in many other common circumstances, e.g., bi-directional prediction for B-frames and EP-frames, deblocking filter, and overlapped block motion compensation (OBMC) [7]. Therefore, in order to extend the applicability of ROPE, cross-correlation must be calculated with manageable complexity.

In fact, this complexity reduction problem has already been addressed in [8], where cross-correlation computation was restricted only to a maximal inter-pixel distance. This approximation is motivated by the fact that two distant pixels are less likely to be averaged in practice. However, this approach still needs to additionally compute and store a substantial number of cross-correlation values in advance. If the cross-correlation of two pixels beyond the maximal distance is needed, it must revert to assuming them uncorrelated, which compromises the estimation accuracy.

The research goal of this work is to seek better approaches to approximate inter-pixel cross-correlation so as to enhance the practical applicability of ROPE. In this paper, two schemes are proposed, stemming from two differing model assumptions. The schemes approximate the cross-correlation as the function of the marginal moments of the two pixels, which are available (see ROPE in [4]). Therefore, they require no additional

0-7803-7750-8/03/$17.00 ©2003 IEEE

storage space. Moreover, as the computation is only performed whenever a specific cross-correlation value is needed, there is no redundant computation for possibly unused cross-correlation values, as would be necessary if one were to recursively estimate the cross-correlation. Simulation results demonstrate the high approximation accuracy of the approaches.

The paper is organized as follows. Section 2 explains the necessity of cross-correlation estimation in applying ROPE with half-pixel prediction. Section 3 details the proposed model-based cross-correlation approximation schemes. Simulation results are summarized in Section 4.

2. ROPE AND SUB-PIXEL PREDICTION

ROPE was originally proposed in [4] as an efficient means to accurately estimate at the encoder the end-to-end distortion. Assuming mean-squared-error (MSE), the end-to-end distortion is:

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

where f_n^i and $\tilde{f}_n^i$ denote the original value and decoder reconstruction value of pixel i in frame n, respectively. Note that due to possible packet loss the decoder reconstruction is viewed at the encoder as a random variable.

As there is no motion compensated prediction in Intra macro-block (MB) coding, we will only focus on the case of Inter mode MB's. For simplicity but without implied loss of generality: (a) We model the channel as a Bernoulli process with packet loss rate *p*. (b) We assume that data of one frame are carried in one packet. Hence, the pixel loss rate equals the packet loss rate. (c) We assume that to conceal a lost frame it is simply replaced by the previous reconstructed frame. As in [4], the moments in (1) are computed recursively by

$$E\{\tilde{f}_n^i\} = (1-p)\cdot(\hat{e}_n^i + E\{\tilde{f}_{n-1}^j\}) + p\cdot E\{\tilde{f}_{n-1}^i\} \tag{2}$$

$$\begin{aligned} E\{(\tilde{f}_n^i)^2\} &= (1-p)\cdot[(\hat{e}_n^i)^2 + 2\hat{e}_n^i E\{\tilde{f}_{n-1}^j\} + E\{(\tilde{f}_{n-1}^j)^2\}] \\ &\quad + p\cdot E\{(\tilde{f}_{n-1}^i)^2\}. \end{aligned} \tag{3}$$

Here, $\hat{e}_n^i$ is the quantized prediction error, and pixel i in frame n is predicted by pixel j in frame *n-1* (given the motion vector).

While H.263 employs half-pixel prediction, in the most recent H.26L standard, sub-pixel prediction has advanced to quarter-pixel accuracy or even better. For simplicity, we restrict the analysis to half-pixel prediction in H.263+ as illustrated in Fig.1 [7]. (Here CTRL is a control parameter with the value of 0 or 1).

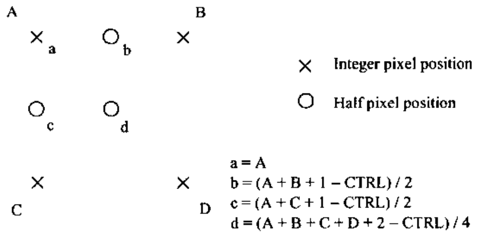

A
B
a
b
c
d
C
D
× Integer pixel position
○ Half pixel position
a = A
b = (A + B + 1 – CTRL) / 2
c = (A + C + 1 – CTRL) / 2
d = (A + B + C + D + 2 – CTRL) / 4

Fig.1 Half-pixel prediction by bilinear interpolation

Assume that pixel *i* in frame *n* is predicted by a half pixel in frame *n-1*, e.g. b as in Fig.1, the 1st and 2nd order moments of b would be

$$E\{\tilde{b}\} = [E\{\tilde{A}\} + E\{\tilde{B}\} + 1 - CTRL]/2 \tag{4}$$

$$\begin{aligned} E\{\tilde{b}^2\} &= [(1-CTRL)^2 + 2\cdot(1-CTRL)\cdot(E\{\tilde{A}\} + E\{\tilde{B}\}) \\ &\quad + E\{\tilde{A}^2\} + E\{\tilde{B}^2\} + 2\cdot E\{\tilde{A}\cdot\tilde{B}\}]/4. \end{aligned} \tag{5}$$

As usual, tilde indicates decoder reconstruction. While (4) can be exactly computed with the already available 1st order moments of the integer pixels A and B, we have to additionally estimate the new quantity $E\{\tilde{A}\cdot\tilde{B}\}$ in (5), i.e. the cross-correlation between A and B. Basically, the presence of cross-correlation is due to *the pixel averaging operation*, which appears not only in sub-pixel prediction, but also elsewhere as explained in Section 1. It is not difficult to see that exact computation of the needed cross-correlation for the current frame may require the availability of all the cross-correlation terms in previous frames, and hence entails too much complexity for practical video coding systems.

3. MODEL-BASED CROSS-CORRELATION APPROXIMATION

The basic idea is to approximate the cross-correlation between two pixels by a function of *the available 1st and 2nd order marginal moments*. Consequently, there will be no additional storage requirements and only minimal additional computational complexity.

We formulate the problem quantitatively and consider two models to capture inter-pixel dependence.

Approximate $E\{XY\}$, given $E\{X\}, E\{Y\}, E\{X^2\}, E\{Y^2\}$. (6)

Model I : $X = a + bY$,
where a, b are unknown constants, $b \geq 0$. (7)

Model II: $X = N + bY$.
b is constant. N is a zero-mean random variable, and is independent of Y. (8)

Given a model assumption, the cross-correlation between X and Y can be expressed in terms of the marginal moments as follows.

For Model I:

$$E\{XY\} = E\{X\}\cdot E\{Y\} + \sigma_X \cdot \sigma_Y. \qquad (9)$$

For Model II:

$$E\{XY\} = \frac{E\{X\}}{E\{Y\}}\cdot E\{Y^2\}, \text{ with } \frac{\sigma_X}{E\{X\}} \ge \frac{\sigma_Y}{E\{Y\}}. \qquad (10)$$

Here σ_X and σ_Y are the standard deviations of X and Y, respectively. Note that in Model I, b is assumed non-negative simply because X, Y are two pixels. The inequality condition in (10) is for notational convenience and simply determines which of the two pixels should play the role of X or Y in the righthand side of (10).

As a practical note, due to error propagation it is worthwhile to apply reasonable bounds on the estimated values. One obvious fact, which can be used to bound the quantities, is that the pixel value should be within the range of 0~255. Cross-correlation can be further bounded by Schwarz inequality as

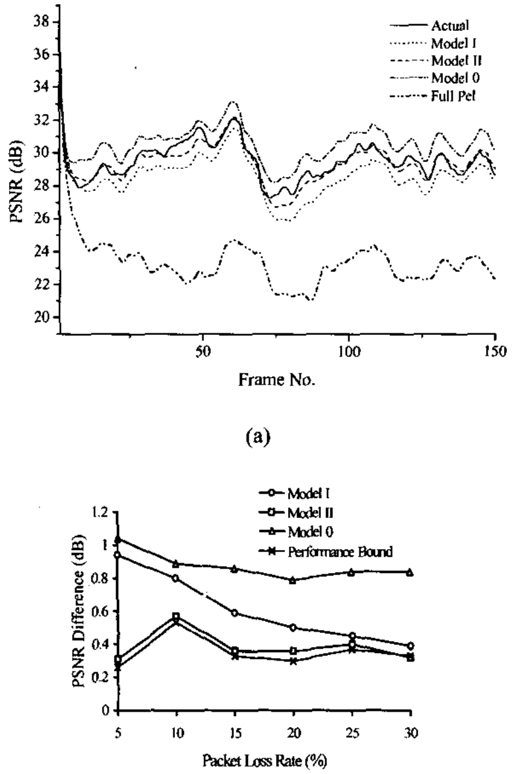
$$E\{XY\} \le \sqrt{E\{X^2\}E\{Y^2\}}. \qquad (11)$$

4. SIMULATION RESULTS

Our simulation setting is based on the UBC H.263+ codec [9]. A sequence is encoded into an H.263 bitstream given the packet loss rate and total bit rate. The bitstream is then decoded with a packet loss pattern that is randomly generated at the prescribed packet loss rate. In the experiments, we use 50 different packet loss patterns. Half-pixel prediction is employed at the encoder.

The proposed two model-based approximation techniques are labeled "Model I" and "Model II", respectively. The method that ignores inter-pixel correlation is denoted by "Model 0", where cross-correlation is assumed to be the product of the means of two pixels. "Full Pel" stands for the method, which approximates half-pixel prediction simply by integer pixel prediction as in [4]. "Actual" in Fig.2a is the real average PSNR result at the decoder. In Fig.2b, we also provide the result of ROPE where the encoder uses integer pixel prediction. This actually demonstrates the best estimation setting for ROPE, and is thus labeled as "Performance Bound".

Fig.2 shows the distortion estimation performance of ROPE given different cross-correlation approximation schemes. In these tests, an MB is coded into Intra mode once per $1/p$ frames, where p is the packet loss rate. In Fig.2a, it is obvious that the proposed "Model II" method has the best end-to-end distortion estimation accuracy among all the tested schemes. In Fig.2b the absolute PSNR difference between the estimated and actual end-to-end distortion is given versus packet loss rate. We can see that the performance of "Model II" approaches the performance bound of ROPE very closely. Also, both proposed methods achieve better estimation accuracy than that of the "Model 0" method.



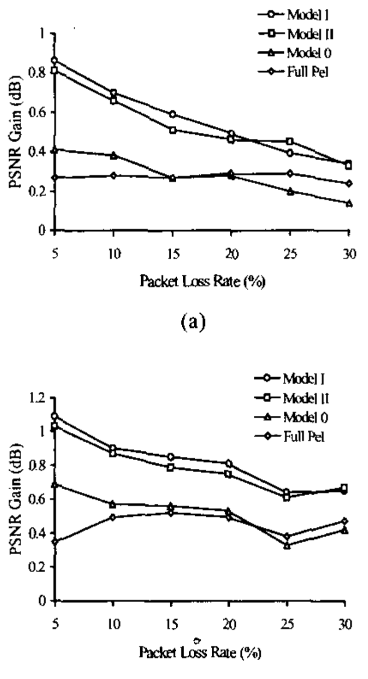
(a)

(b)

Fig.2 Estimation performance comparison. Periodic Intra updating, "foreman", QCIF, 30f/s, 200kb/s, 1st 150 frames. (a) p=5%.

As is well known, performance gains can be achieved by applying half-pixel prediction. In Fig.3, we compare the performance improvement achieved by different cross-correlation approximation schemes. The curves provide the gains of ROPE with half-pixel motion compensation relative to ROPE with integer pixel motion compensation. It is easy to see that both proposed approximation schemes consistently achieve better performance gains than the other two methods. Note that the result of "Full

Pel" is the result of ROPE as proposed in [4]. Hence, we conclude that accurate approximation of cross-correlation guarantees the performance improvement of applying ROPE with half-pixel prediction. More generally, the practical applicability of ROPE is significantly enhanced by the proposed approximation techniques.



(a)

(b)

Fig.3 Performance improvement comparison. R-D optimized Intra/Inter coding mode selection, QCIF, 30f/s, 1st 150 frames. (a) "foreman", 200kb/s. (b) "miss_am", 100kb/s.

5. CONCLUSION

In spite of the remarkable performance of ROPE on end-to-end distortion estimation, its applicability in practical video coding systems is limited by the open problem of cross-correlation estimation, which is encountered in many common circumstances of video coding involving pixel averaging operations. In this paper, we propose two model-based schemes to approximate the cross-correlation with a function of the marginal moments (available quantities). The proposed methods are efficient in that there is no additional storage requirements and minimal additional computational cost. More notably, the end-to-end distortion estimation accuracy with the proposed approximation is strikingly close to the ROPE performance bound.

REFERENCES

[1] A. Ortega and K. Ramchandran, "Rate-Distortion Methods for Image and Video Compression," *IEEE Signal Processing. Magzine*, Nov. 1998, pp. 23-50.

[2] G. J. Sullivan and T. Wiegand "Rate-Distortion Optimization for Video Compression," *IEEE Signal Processing Magzine*, Nov. 1998, pp. 74-90.

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

[4] 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, no.6, June 2000, pp. 966-976.

[5] A. Reibman, "Optimizing Multiple Description Video Coders in a Packet Loss Environment," *Proc. of PVW 2002.*

[6] H. Yang, R. Zhang and K. Rose, "Drift Management and Adaptive Bit Rate Allocation in Scalable Video Coding," *Proc. of ICIP 2002.*

[7] ITU-T Recommendation H.263, "Video Coding for Low Bitrate Communication," 1998.

[8] K. Stuhlmuller, "Modeling and Optimization of Video Transmission Systems," *Shaker Verlag*, 2000, pp.46.

[9] H.263+ codec, http://spmg.ece.ubc.ca/