Rate-Distortion Modeling for Wavelet Video Coders

Mingshi Wang
Department of Electrical and
Computer Engineering
University of California, Davis
Email: verwang@ece.ucdavis.edu

Mihaela van der Schaar
Department of Electrical and
Computer Engineering
University of California, Davis
Email: mvanderschaar@ece.ucdavis.edu

***Abstract*— Based on our statistical investigation of a typical 3D (t+2D/2D+t) wavelet codec, we present a unified mathematical model to describe its rate-distortion (RD) behavior. In this work we assess the quantization distortion of the reconstructed video frames by tracking the quantization noise along the 3D wavelet decomposition trees. The minimum bit rate to achieve this distortion is estimated by the unconditional entropy of the quantizer. Experimental results show that the model captures sequence characteristics accurately and reveals the relationship between wavelet decomposition levels and the overall RD performance.**

I. Introduction

In recent years, the demand for media-rich applications over wired and wireless IP networks has grown significantly. This requires video coders to support a large range of scalability. Standard coders such as H.26x and MPEG-x have very limited scalability due to the close loop prediction structures. As an improvement, wavelet-based coders with open loop temporal filtering structure support SNR, spatial and temporal scalability [16]. Therefore it is useful to find concrete analytical RD models for wavelet-based coders to guide the real-time adaptation of video bitstream sto instantly varying network condition.

Generally there are two types of methodologies for RD analysis. The first is the empirical approach [1], [2], [4] where experimental RD data is fitted to derive functional expressions. Though this approach is easy to carry out, it neither considers explicitly the characteristics of the input sequence nor models the RD behavior of the coding algorithm. Therefore, it can not be directly applied to different video sequences or coding structures. The second is the analytical approach based on traditional RD theory [9]–[11]. It is known that explicit functions for the RD performance are only available for stationary Gaussian processes [17]. Based on this assumption, a theoretical RD model for simple transform video coders has been derived by Hang et. al. [6], [7]. In Hang's work, the bit rate was estimated under the assumption of a simple Gaussian source and small quantization steps. Therefore, the estimation is not accurate for low bit rates and cannot be generalized to more sophisticated transform coders. In Girod's work [3], [5], the propagation of power spectral density of prediction error was derived for the close loop MC structure of a coder. While this approach presents a model for analyzing the quantization distortion for a close loop codec, it is limited by the simplified assumptions of the prediction filtering process and hence cannot describe the RD behavior of open-loop wavelet video codecs accurately.

He's model [8] lies between analytical and empirical methodologies. The rate curves are first decomposed into pseudo-coding bit rates for the zero and nonzero coefficients and then each part is adapted empirically by the training data. This model successfully captures the input source characteristics but it only considers the effect of quantization steps on the RD performance of still images.

In this paper, we derive an analytical model of a typical wavelet video codec by analyzing the propagation of quantization distortion within the 3D wavelet tree, and the bit rate is approximated by the unconditional quantizer entropy. Our objective is to quantify the relationship between wavelet decomposition structure, bit rate and distortion and therefore guide the multimedia delivery and access. The contributions of this paper are as follows: (1) the model developed in this paper is shown to be applicable to various wavelet coders with different MCTF structures; (2) the model can be used to enable efficient adaptation to live conditions and therefore can be employed in both serve-driven and receiver-driven streaming paradigms. The paper is organized as follows. Part II gives the theoretical derivation of the RD model. Part III demonstrates the accuracy of the model by adapting the model to some experimental data and Part IV concludes this paper.

II. RD Modeling for Wavelet-Based Video Codecs

In this section, we present our framework for modeling the RD performance of wavelet-based video coders. We start by summarizing several results already derived for image data (spatial domain), and then extend this analysis to the temporal domain.

A. Quantization Distortion in the Spatial Domain

Both goodness-of-fit tests and theoretical analysis proved that the DWT subband coefficients approximately obey a Laplacian distribution. Moreover, this result also holds for the transformed coefficients of motion-compensated (residual) video frames [12]–[14]. Hence, the quantization distortion of a Laplacian variable with variance σ for an N level uniform quantizer with deadzone g_0 is [6]:

$$\varepsilon(g_0,\Delta,\sigma) = -\rho_1\big(g_0-\frac{\Delta}{2}+\frac{1}{\lambda}\big)^2+\frac{2}{\lambda^2}-\frac{\rho_1\rho\Delta^2(1-\rho^{N-1})}{(1-\rho)^2} \quad (1)$$

where Δ is the quantization step size, $\rho \triangleq e^{-\lambda\Delta}$, $\rho_1 \triangleq e^{-\lambda(g_0-\Delta/2)}$, and $\lambda \triangleq \frac{\sqrt{2}}{\sigma}$. In order to minimize the quantization distortion, the reconstruction level is chosen to be the centroid of each quantization cell. The corresponding quantizer entropy is:

$$\begin{aligned} h(g_0,\Delta,\sigma) = &-(1-\rho_1)\log(1-\rho_1) - \rho_1\log\rho_1 \\ &- \rho_1(1-\rho^{N-1})\log(1-\rho) + \rho_1 \\ &- \frac{\rho_1\rho(1-\rho^{N-1})}{1-\rho}\log\rho \qquad \text{(bits/sample)} \end{aligned} \tag{2}$$

For a J-level spatial wavelet transform, there are $3J+1$ subbands. Let the quantization distortion within the k-th subband in level j be $\varepsilon_j^{(k)}$, $k = 1,2,3$. Due to the orthonormality of DWT, the signal variance E_0 in the original frame is distributed into the k-th subband with approximately fixed ratio β_k, $k = 0,1,2,3$, where the subscript $k = 0,1,2,3$ indicates the LL, LH, HL and HH subbands. This leads to the following result: $\varepsilon_j^{(k)} = \varepsilon(g_0,\Delta,\beta_0^{j-1}\beta_k E_0)$. Therefore, the quantization distortion in the reconstructed frame after inverse DWT is [14], [17]:

$$\begin{aligned} d(g_0,\Delta,J,E_0) = &\, 4^{-J}\varepsilon(g_0,\Delta,\beta_0^J E_0) \\ &+ \sum_{j=1}^{J}\sum_{k=1}^{3} 4^{-j}\varepsilon(g_0,\Delta,\beta_0^{j-1}\beta_k E_0) \end{aligned} \tag{3}$$

and the average entropy is:

$$\begin{aligned} \mathbf{h}(g_0,\Delta,J,E_0) = &\, 4^{-J}h(g_0,\Delta,\beta_0^J E_0) \\ &+ \sum_{j=1}^{J}\sum_{k=1}^{3} 4^{-j}h(g_0,\Delta,\beta_0^{j-1}\beta_k E_0) \\ &\qquad \text{(bits/sample)} \end{aligned} \tag{4}$$

Equation (3) and (4) indicate that the distortion and rate of a reconstructed frame is a function of the spatial decomposition levels, the quantizer parameters and also the average signal variance of the source (image).

B. Effect of Motion-Compensation Temporal Filtering on Quantization Noise Propagation

Motion-compensation temporal filtering (MCTF) can be employed either before or after spatial wavelet decomposition. These two schemes are called t+2D and 2D+t MCTF, respectively. In this section, we analyze the average frame distortion at different temporal levels for t+2D MCTF. However, using a similar analysis as the one outlines below, the RD derivation can be easily extended to the 2D+t MCTF structure.

Current wavelet video coders typically use the Haar filter and 5/3 filter for temporal filtering. However, other filters, such as the longer 9/7 filters can also be used to exploit the time dependency between successive video frames and hence improve the RD performance. We will show in this section that the RD model for the various temporal filters can be derived in the same manner.

First, we analyze a wavelet video coder using Haar temporal filtering. In a temporal filtering structure with T levels, there are 2^T frames in one Group Of Frames (GOF). Assume approximately constant signal variance E_0 within one GOF and let A and B stand for the even and odd frames, respectively, in the lifting structure [15]. In motion estimation process, the pixels can be classified into three types: connected, unconnected and multiple connected. For most video sequences, frame A has only a small portion of pixels that do not have a correspondence in the reference frame, and are thus intra-coded [18]. To simplify the analysis, let us assume all the pixels in frame A fall into two categories: connected and multiple connected. Let r_c be the ratio of connected pixels, r_u be the ratio of unconnected pixels and r_m be the multiple connected pixels. Due to the above assumptions we have $r_c + r_m = 1$ and $r_u = r_m$ [18]. Denote the low and high pass frames in the k-th temporal level as $I^{(k)}$ and $H^{(k)}$, where the superscript denotes the temporal level, and

$$\begin{aligned} &E[I^{(k)}(m,n,t)\mathrm{MC}(I^{(k)}(m,n,t+2^k l))] = \\ &\quad E[\mathrm{IMC}(I^{(k)}(m,n,t))I^{(k)}(m,n,t+2^k l)] \cong r(l) \cong r \end{aligned} \tag{5}$$

From the lifting structures, the variances of the two types of pixels - $I_c^{(k)}$ and $I_u^{(k)}$ can be determined as:

$$\begin{aligned} &E\left[\left|I_c^{(k)}(m,n,t)\right|^2\right] = \\ &\frac{1}{2}\Big[E\left[\left|I_c^{(k-1)}(m,n,t)\right|^2\right] + E\left[\left|\mathrm{MC}(I_c^{(k-1)}(m,n,t+2^k))\right|^2\right] \\ &+ 2E\left[I_c^{(k-1)}(m,n,t)\mathrm{MC}(I_c^{(k-1)}(m,n,t+2^k))\right]\Big] \end{aligned}$$

for connected pixels, and $E\left[\left|I_u^{(k)}\right|^2\right] = 2E\left[\left|I_u^{(k-1)}\right|^2\right]$ for unconnected pixels. Note $E\left[\left|I^{(0)}\right|^2\right] = E\left[A_c^2\right] \cong E\left[B_c^2\right] \cong E_0$, according to the assumption (equation (5)) we have $E\left[\left|I_c^{(k)}\right|^2\right] = (1+r)E\left[\left|I_c^{(k-1)}\right|^2\right]$. Taking the average gives:

$$\begin{aligned} E\left[\left|I^{(k)}\right|^2\right] &= r_u 2E\left[\left|I^{(k-1)}\right|^2\right] + (1-r_u)(1+r)E\left[\left|I^{(k-1)}\right|^2\right] \\ &= [2r_u + (1+r)(1-r_u)]E\left[\left|I^{(k-1)}\right|^2\right] \\ &= [2r_u + (1+r)(1-r_u)]^k E_0 \end{aligned} \tag{6a}$$

Similarly, the signal power in the high pass frame is:

$$E\left[\left|H^{(k)}\right|^2\right] = (1-r)[2r_u + (1+r)(1-r_u)]^{k-1}E_0 \tag{6b}$$

Equation (6) shows that the compaction in the signal energy increases exponentially towards the lower temporal subband.

The propagation of quantization noise along the wavelet tree has been studied intensively by Rusert et. al. [18], and their result shows that the average distortion for a pair of A and B frames can be expressed as:

$$\begin{aligned} d_I^{(0)} &= \frac{1}{2}(d_A + d_B) \\ &= \frac{1}{2}d_I^{(1)} + (\frac{3}{4} - \frac{r_c}{4})d_H^{(1)} \end{aligned} \tag{7}$$

where $d_I^{(k)}$ and $d_H^{(k)}$ denote the distortion in frame $I^{(k)}$ and $H^{(k)}$ Taking average on both sides of equation (7), we derive

the average frame distortion at temporal levels 0 and 1:

$$\bar{d}_I^{(0)} = (\frac{3}{4} - \frac{\bar{r}_c}{4})d_H^{(1)} + \frac{1}{2}\bar{d}_I^{(1)} \tag{8}$$

where the bar indicates the average value. Generally, the average distortion of I frames in the k-th temporal level is given by:

$$\bar{d}_I^{(k-1)} = (\frac{3}{4} - \frac{\bar{r}_c(k)}{4})d_H^{(k)} + \frac{1}{2}\bar{d}_I^{(k)} \tag{9}$$

which leads to

$$\bar{d}_I^{(0)} = \sum_{k=1}^{T}(\frac{3}{4} - \frac{\bar{r}_c(k)}{4})(\frac{1}{2})^{k-1}d_H^{(k)} + (\frac{1}{2})^T\bar{d}_I^{(T)} \tag{10}$$

It should be noted that $d_I^{(k)}$ and $d_H^{(k)}$ are given by equation (3) with the signal variance replaced by equation (6). The average entropy within one GOF is calculated to approximate the output bit rate of the entropy encoder:

$$\begin{aligned}\mathcal{H} &= 2^{-T}\mathbf{h}(\Delta/2^{T-1}, g_0/2^{T-1},\\ &\quad J, [2r_u + (1+r)(1-r_u)]^T E_0)\\ &\quad + \sum_{k=1}^{T} 2^{-k}\mathbf{h}(\Delta/2^{k-1}, g_0/2^{k-1},\\ &\quad J, (1-r)[2r_u + (1+r)(1-r_u)]^{k-1}E_0)\end{aligned} \tag{11}$$

(bits/sample)

Therefore equation (10) and (11) approximately describe the RD behavior for the Haar filter case.

The above derivation is only valid under the assumption of accurate inversibility of motion estimation. However, this assumption does not hold for cases such as sub-pixel interpolation [16], [19], [20]. On the other hand, the lifting structures for 5/3 and 9/7 filters are much more complicated than that for the Haar filter, which complicates the tracktability of the quantization error along the temporal wavelet tree. However, equation (9) suggests that we can always find a linear relationship between the average frame distortions within adjacent temporal levels:

$$\bar{d}_I^{(k)} = A^{(k+1)}\bar{d}_I^{(k+1)} + B^{(k+1)}d_H^{(k+1)} \tag{12}$$

This tells us that the average distortion for the original video sequences can be expressed as:

$$\bar{d}_I^{(0)} = \sum_{k=1}^{T} B^{(k)} \prod_{j=1}^{k-1} A^{(j)} d_H^{(k)} + \prod_{j=1}^{T} A^{(j)}\bar{d}_I^{(T)} \tag{13}$$

Hence, equation (10) is a special form of equation (13). It is seen that the parameters $A^{(k)}$ and $B^{(k)}$ are determined by the lifting structures and the method of motion estimation, with $B^{(k)} = (\frac{3}{4} - \frac{\bar{r}_c(k)}{4},)$, $A^{(k)} = \frac{1}{2}$ for Haar filter in the simplest case.

Similar to the Haar filter case, we can derive the signal variance distribution for 5/3 and 9/7 filters from their respective analysis transfer functions. But due to space limitation, we only reproduce our result for the 5/3 filters, while omitting its derivation:

$$\begin{aligned}&E\left[\left|H^{(k)}(m,n)\right|^2\right] \cong\\ &\quad (\frac{1}{4}r(2) - r(2) + \frac{3}{4})\times\\ &\quad (\frac{1}{16}r(4) - \frac{1}{4}r(3) - \frac{1}{2}r(2) + \frac{5}{4}r(1) + \frac{23}{16})^{k-1}E_0\end{aligned} \tag{14a}$$

$$\begin{aligned}&E\left[\left|I^{(k)}\right|^2\right] \cong\\ &(\frac{1}{16}r(4) - \frac{1}{4}r(3) - \frac{1}{2}r(2) + \frac{5}{4}r(1) + \frac{23}{16})^{k}E_0\end{aligned} \tag{14b}$$

It is also easy to evaluate the average entropy since it has the same form as equation (11). The only change is to use appropriate expressions for signal variances.

III. Experimental Results

With the model developed in section II, we optimized the parameters in the model with respect to our experimental data obtained with the 3D ESCOT codec [16]. Fig. 1 presents the rate distortion behavior for two representative video sequences: Coastguard and Akiyo. In this scenario the temporal decomposition level is set to 4, and the spatial decomposition level is varied from 1 to 3. The codec is set to t+2D MCTF mode using 5/3 wavelets and the frame rate is set to 30 Hz. The wavelet coefficients of each subband is compressed to an embedded bitstream using fractional bitplane coding. In the experiment we truncated the bitstream at the following

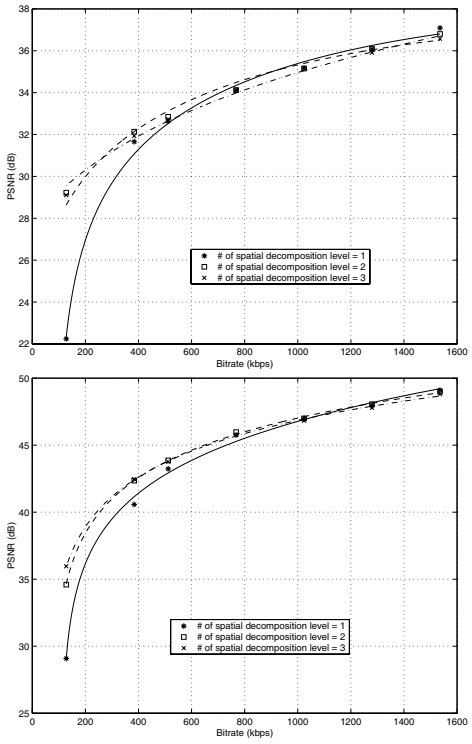

Fig. 1. Rate distortion curve of two sequences at 3 spatial decomposition levels with 4 the temporal levels. From top to bottom: "coastguard", "Akiyo".

bit rates: 128, 384, 512, 768, 1024, 1280, 1536 (kbps). The model is fitted by choosing three experimental data points at 128, 768 and 1536 (kbps). The theoretical curve is drawn with solid lines and experimental data is shown with symbols. These results indicate that the model successfully captures the characteristics of the video sequences with the theoretical curves passing through most of the experimental data points. At low bit rates higher spatial decomposition level results in considerably lower distortion due to the fact that the energies of each frame are well compacted in the spatial domain and hence, less bits are needed to achieve a better PSNR. But at higher bit rates, a higher spatial decomposition level is not always a better solution.

In another scenario shown by Fig. 2, we fixed the spatial decomposition level to 3, while changing the temporal filtering level from 2 to 4. Again the model is adapted using three training data points and it successfully predicts the other experimental data points.

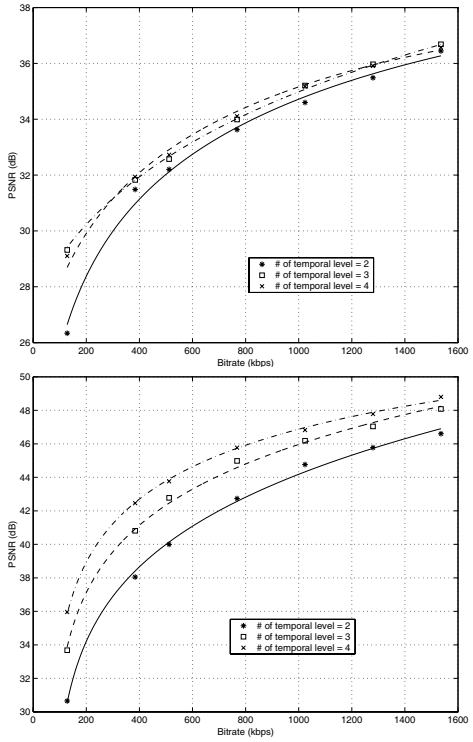

Fig. 2. Rate distortion curve of two sequences at 3 different temporal decomposition levels with 3 sptial decomposition levels. ¿From top to bottom:“coastguard”, “Akiyo”.

IV. Conclusions

In this paper we developed an analytical model for a wavelet video codec. The model estimates the RD curve accurately and efficiently. The average bit rate used to encoding the video sequences are estimated from the unconditional quantization entropy. Since most codecs use some sort of conditioning and contexts, it will be an interesting topic to investigate it further.

References

[1] W. Ding and B. Liu, “Rate control of MPEG video coding and recording by rate-quantization modeling,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 12-20, Feb., 1996.

[2] T. Chiang and Y. Zhang, “A new control scheme using quadratic rate distortion model,” *IEEE Trans. CirCuits Syst. Video Technol.*, vol. 7, pp.246-250, Feb., 1997.

[3] B. Girod, “The efficiency of motion-compensating prediction for hybrid coding of video sequences,” *IEEE J. Select. Areas Commun.*, vol. SAC-5, pp. 1140-1154, Aug., 1987.

[4] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, “Analysis of video transmission over lossy channels,” *IEEE J. Select. Areas Commun.*, vol. 18, pp.1012-1032, June, 2000.

[5] B. Girod, “Motion-compensating prediction with fractional-pel accuracy,” *IEEE Trans. Commun.*, vol. 41, pp.604-612, Apr., 1993.

[6] H. Hang and J. Chen, “Source model for transform video coder and its application–Part I: Fundamental theory,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 287-298, Apr., 1997.

[7] J. Chen and H. Hang, “Source model for transform video coder and its application–Part II: Variable frame rate coding,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 299-311, Apr., 1997.

[8] Z. He and S. K. Mitra, “A unified rate-distortion analysis framework for transform coding,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 1221-1236, Dec., 2001.

[9] J. B. O Neal, JR. and T. Raj Natarajan, “Coding isotropic images,” *IEEE Trans. Inform. Theory*, vol. IT-23, pp.697-707, Nov., 1997.

[10] D. J. Sakrison, “The rate distortion function of a Gaussian process with a weighted square error criterion,” *IEEE Trans. Inform. Theory*, vol. 14, pp.506-508, May, 1968.

[11] D. J. Sakrison, “A geometric treatment of the source encoding of a Gaussian random variable,” *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 481-486, May, 1968.

[12] F. Bellifemine, A. Capellino, A. Chimienti, R. Picco, and R. Ponti, “Statistical analysis of the 2-D coefficients of the differential signal for images,” *Signal Processing, Image Commun.*, vol. 4, pp. 477-488, 1992.

[13] E. Y. Lam and J. W. Goodman, “A mathematical analysis of the DCT coefficient distributions for images,” *IEEE Trans. Image Processing*, vol. 9, pp. 1661-1666, Oct., 2000.

[14] S. G. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 674-693, July, 1989.

[15] I. Daubechies and W. Sweldens, “Factorization wavelet transforms into lifting steps,” *J. Fourier Anal. Appl.*, vol. 4, pp. 247-269, 1998.

[16] J. Xu, Z. Xiong, S. Li, and Y. Zhang, “Three-dimensional embedded subband coding with optimized truncation (3-D ESCOT),” *Appl. Comun. Harmonic Anal.*, vol. 10, pp. 290-315, 2001.

[17] D. S. Taubman and M. W. Marcellin, *JPEG 2000-Image Compression Fundamentals, Standards and Practice*, Kluwer Academic Publishers, 2002.

[18] T. Rusert, K. Hanke, and J. Ohm, “Transition filtering and optimization quantization in interframe wavelet video coding,” *VCIP, Proc. SPIE*, vol. 5150, pp. 682-693, 2003.

[19] J. Ohm, M. van der Schaar, J. W. Woods, “Interframe wavelet coding-motion picture representation for universal scalability,” *Image Communication*, Special issue on digital cinema, 2004.

[20] J. R. Ohm, *Multimedia Communication Technology*, Springer, 2004.