Operational Rate-Distortion Modeling for Wavelet Video Coders

Mingshi Wang and Mihaela van der Schaar, *Senior Member, IEEE*

***Abstract*—Based on our statistical investigation of a typical three-dimensional (3-D) wavelet codec, we present a unified mathematical model to describe its operational rate-distortion (RD) behavior. The quantization distortion of the reconstructed video frames is assessed by tracking the quantization noise along the 3-D wavelet decomposition trees. The coding bit-rate is estimated for a class of embedded video coders. Experimental results show that the model captures sequence characteristics accurately and reveals the relationship between wavelet decomposition levels and the overall RD performance. After being trained with offline RD data, the model enables accurate prediction of real RD performance of video codecs and therefore can enable optimal RD adaptation of the encoding parameters according to various network conditions.**

***Index Terms*—Context coding, discrete wavelet transform (DWT), entropy, rate-distortion function.**

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 such as the spatial, temporal and signal-to-noise ratio (SNR) scalability. Standard coders such as H.26x and MPEG-x have limited scalability due to the closed-loop prediction structures. On the other hand, wavelet video coding gained much research interest in the past years, since it not only enables a wide range of scalabilities but also provides high compression performance, comparable to state-of-the-art single layer coders such as MPEG-4 AVC. [23]. Therefore, it is useful to find concrete analytical rate-distortion (RD) models for wavelet-based coders to guide the real-time adaptation of video bitstreams to instantly varying network conditions.

A. Review of Existing RD Modeling Work

Generally, there are two types of methodologies for RD analysis. The first is the empirical approach [9], [10], [20], where experimental RD data are fitted to derive functional expressions. This approach is advantageous because the RD models can be easily computed, but it has the drawback that it does not explicitly consider the input sequence characteristics or the encoding structure and parameters, and hence, the obtained RD performance cannot be generalized. We will resort to the second approach, which is the analytical approach based on traditional RD theory [2], [3], [13].

Finding accurate mathematical RD models for a given random process is generally difficult, and exact derivations are only available for several very simple sources under appropriate fidelity critera. For example, Sakrison [2], [3] calculated a parametrical RD function for a Gaussian process using a weighted square error criterion; Gerrish [1] showed that the Shannon lower bound is attainable if and only if a process can be decomposed into two statistically independent processes with one of them being a white Gaussian process; Berger [4] derived an explicit RD formula for both time-discrete and time-continuous Wiener processes. For most general cases, only upper and lower bounds can be derived for the RD functions of a random process. Note that such bounds are useful for the purpose of RD estimation only when they are tight enough.

Based on these results, a theoretical RD model for simple transform-based video coders has been derived by Hang *et al.* [11], [12]. In this research, the bit rate was estimated under the assumption of a simple Gaussian source model and small quantization steps. Therefore, the estimation is not accurate in the region of low bit rate and cannot be generalized to more sophisticated transform-based coders. In Girod's work [5], [8], the propagation of power spectral density of prediction error was derived for the closed-loop motion-compensated (MC) prediction 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. Mallat *et al.* [17] provided a thorough analysis of RD performance of wavelet transform coding in the low bit-rate region and presented an accurate model for still-image coding. These results revealed that the RD function takes very different forms for low and high bit rates. He's model [26], [28] revealed a linear relationship between the coding bit rates and the percentage of zero-quantized coefficients (also called ρ-domain analysis). In He's model, the rate curves are first decomposed into pseudocoding bit rates for the zero and nonzero coefficients, and then each part is adapted empirically by the training data. This model captures the input source characteristics successfully and leads to an easy approach for estimating the bit rates for video coding [28]. Dai *et al.* [34] combined the distortion calculation in Mallat's work [17] with the bit-rate estimation in He's work [26], [28] and arrived at a closed-form operational RD model. While these models [17], [26], [28], [34] describe the operational RD behavior for a wide range of transform-based coders successfully,

Manuscript received December 4, 2004; revised October 12, 2005. This work was supported in part by the National Science Foundation under a CAREER award, Intel IT research, and UC Micro. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Trac D. Tran.

M. Wang is with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA (e-mail: verwang@ucdavis.edu).

M. van der Schaar is with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095-1594 USA (e-mail: mihaela@ee.ucla.edu).

Digital Object Identifier 10.1109/TSP.2006.879273

they are expressed either in terms of quantization step sizes or value of percentage of zero coefficients after quantization, i.e., the value of ρ. Since there exists a one-to-one mapping between the quantization step size and ρ, the rate control algorithm based on these models is restricted to the adjustment of quantization step sizes in order to meet certain bit-rate requirements. However, other coding parameters besides the quantization step sizes, such as the number of wavelet decomposition levels, the choice of wavelet filters, etc., have a significant impact on the RD behavior of the source coder.

B. The Statistical Properties of Wavelet Coefficients and Our RD Model

We will derive an analytical model for the RD behavior of a typical t+2D wavelet video coding system [35]. The distortion is calculated by tracking the propagation of quantization noise along the three-dimensional (3-D) wavelet decomposition trees. The estimation of the coding bit rate requires an accurate and simple joint statistical model of the wavelet-domain coefficients. There have been various investigations on the statistical distributions of the transform-domain coefficients of natural images and the generalized Gaussian distribution is demonstrated to describe the marginal distribution of the coefficients very accurately [7], [19]. Even though the correlation of transform-domain coefficients is nearly zero, which demonstrates the decorrelation property of orthogonal transforms such as discrete cosine transform and discrete wavelet transform (DWT), dependencies remain among these coefficients both across and within different subbands [31]. Of particular interest is the joint statistical model of DWT coefficients, which can be assorted into two classes: intrasubband [14], [18], [25], [33], and intersubband dependencies [16], [22], [24]. The intrasubband dependencies are characterized by doubly stochastic processes, where the state variables are closely correlated among neighboring coefficients. The dependencies across scales (also called persistency) are modeled by a hidden Markov process. Conditioned on the state variables in both cases, the wavelet coefficients are independent identically distributed Gaussian random variables. Given appropriate statistical distribution of the underlying state variables, the marginal distribution of the wavelet coefficients can be shown to be generalized Gaussian, which explains the statistical models in [7] and [19].

In view of the inter- and intrascale dependency, the encoding bit rate can be greatly reduced by context-based entropy coding [31]. In coding of wavelet coefficients such as the embedded zerotree and the set partitioning in hierarchical trees (SPIHT) [27], dependencies are exploited from the quadtree which relates the parent and child coefficients from subbands of successive scales in the same orientation. One drawback of zerotree-based algorithms is that they do not facilitate resolution scalability due to the fact that a quadtree typically spans different scales. Therefore, subbands belonging to different scales have to be encoded independently to facilitate resolution scalability. Moreover, it was found that the coding penalty resulting from not exploiting the interscale dependencies is minimal [25]. This was verified by recently developed codecs such as EBCOT [21], which demonstrated compression performance comparable to that of SPIHT. Hence, we will restrict our analysis to independent subband context coding.

C. Objective and Organization of this Paper

Our objective is to quantify the relationship between the spatiotemporal wavelet decomposition structure, bit rate, and distortion. An RD model is developed that is shown to be applicable to various wavelet video coders with different motion-compensated temporal filtering (MCTF) structures. Section II summarizes previous results on the modeling of wavelet coefficients and presents bit-rate estimation for the coding of a single frame with a context-based entropy coder. In Section III, we focus on the temporal wavelet decomposition tree of a video sequence and derive a recursive expression for the average quantization distortion within each temporal level. Section IV summarizes the RD modeling procedure. Section V demonstrates the accuracy of the model by adapting it to some experimental data. Section VI concludes this paper.

II. The RD Model of a Single Frame

MCTF decorrelates the video sequences in the temporal axis and can be employed either before or after two-dimensional (2-D) spatial wavelet decomposition. These two schemes are called t+2D and 2D+t wavelet decomposition, respectively [35]. In our case, we will focus on the t + 2D scheme, where the 2-D spatial decomposition is performed within all the temporal high-frequency (H) and low-frequency (L) frames of the MCTF decomposition, resulting in multiple 3-D subbands.

We first summarize the statistical properties of wavelet coefficients for both L and H frames, then derive the bit-rate estimation and quantization distortion for a subband in a t + 2D wavelet decomposition tree. Finally, we analyze the problem of optimum bit-rate allocation and the corresponding operational rate-distortion function.

A. Review of Statistical Properties for Wavelet Image Coding

Natural images can be viewed as 2-D random fields that are characterized by singularities such as edges and ridges. The DWT decomposes an image into a multiresolution representation (or a set of transform coefficients in several scales) [6]. The smooth regions in the original image are represented by large coefficients in the coarsest resolution, while edges and ridges are represented by few clustered large coefficients (detail signals) in different resolutions. This property of the DWT leads to a high-peaked heavy-tailed non-Gaussian wavelet-coefficient distribution within each subband. Despite the fact that wavelet coefficients are approximately decorrelated, there remains a certain degree of dependency between them. To capture the non-Gaussianity and the remaining intrascale dependency, we model the wavelet coefficients as a doubly stochastic process [14], [18] parameterized by Θ, which is itself following a Laplacian distribution

$$\begin{aligned} X &\sim N(0,\theta) \\ \Theta &\sim p(\theta) = \frac{1}{\sigma^2}\mathrm{e}^{-\frac{1}{\sigma^2}\theta}. \end{aligned} \tag{1}$$

The marginal distribution of the wavelet coefficients is

$$p(x) = \int p(x|\theta)p(\theta)\mathrm{d}\theta = \frac{1}{\sqrt{2}\sigma}\exp\left\{-\sqrt{2}\frac{|x|}{\sigma}\right\}. \tag{2}$$

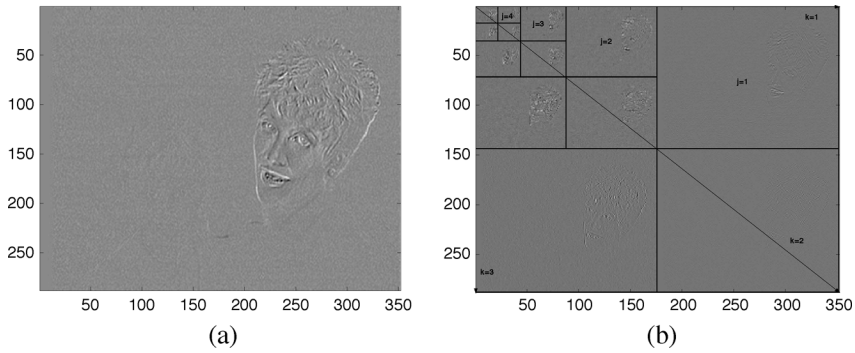


Fig. 1. The H frame before and after spatial wavelet decomposition. (a) The original H frame and (b) the H frame after a four-level 2-D spatial decomposition.que

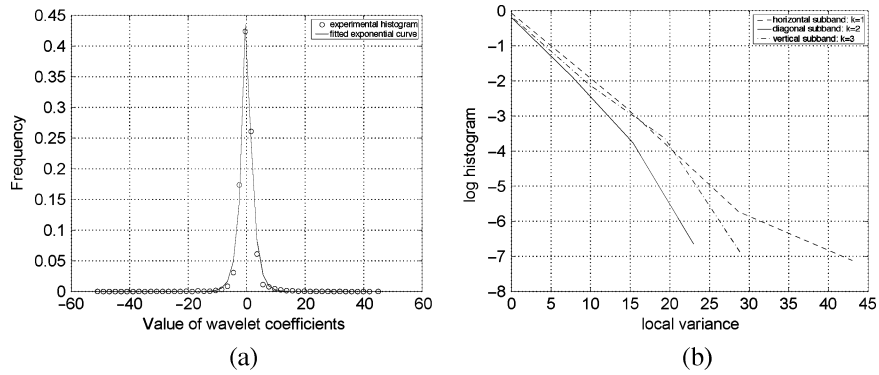


Fig. 2. The statistical properties of the H frame. (a) The normalized histogram of the horizontal subband at scale $j = 3$; the symbols represent the experimental histogram and the solid line is the fitted Laplacian curve. (b) The log-histogram of the local state variable Θ for the three subbands (horizontal, vertical, and diagonal) at scale $j = 3$. The total number of spatial decomposition is four.

The last equation indicates that the wavelet coefficient within each subband obeys the Laplacian distribution, which is consistent with the results in [7] and [19]. Moreover, this doubly stochastic model is found to describe accurately the H frames generated by temporal filtering in a wavelet video codec. For a J-scale 2-D spatial-domain DWT, there are $3J+1$ subbands. Let $j(1 \leqslant j \leqslant J)$ denote the scale number and $k = 1, 2, 3$ represent the horizontal, diagonal, and vertical orientations, respectively.

Fig. 1(a) gives an example of a H frame from the Mother and Daughter sequence, and Fig. 1(b) is the H frame after a four-level 2-D spatial decomposition by Daubechies 5/3 filters. The normalized histogram of the horizontal subband at scale $j = 3$ is demonstrated by the disjoined points in Fig. 2(a). It is seen that the histogram fits closely with the function of (2) plotted by the solid line. By following a similar method developed in [25], we plot in Fig. 2(b) the log-histogram of the local variance Θ for the three subbands (horizontal, vertical and diagonal) at scale $j = 3$. The local variance of a pixel is estimated from the eight adjacent wavelet coefficients, and all the log-histograms appear to be linear functions of the local variance Θ, which verifies our assumption given in (1).

The doubly stochastic model of (1) also leads to the Markov property of wavelet coefficients. Denote the neighbor of the current wavelet coefficient X by $\mathcal{N}X$. Then, according to the Markov property of the hidden state, X is conditionally independent of $\mathcal{N}X$ when its state variable is given, i.e., the following sequence forms a Markov chain [25]:

$$X \longrightarrow \Theta \longrightarrow \mathcal{N}X. \tag{3}$$

The interscale dependencies are best described by the mixture Gaussian process of [22]. Let σ_j^2 be the signal variance in scale $j, 1 \leqslant j \leqslant J$. From the Markov-state model [22, (7)], σ_j^2 can be expressed by

$$\sigma_j^2 = \mu^{2j}\sigma_J^2 + (1-\mu^{2j})\sigma_w^2 = \mu^{2j}\left(\sigma_J^2 - \sigma_w^2\right) + \sigma_w^2 \tag{4}$$

where σ_J^2 is the signal variance of the coarsest resolution, $|\mu| < 1$ is a constant, and σ_w^2 is the variance of the white Gaussian noise which drives the autoregressive state equation. When $\sigma_w^2 \ll \sigma_J^2$, (4) implies an approximately exponential decay of the signal variance from the coarsest to the finest scales, which is indeed verified by the results of Fig. 3. We applied the Daubechies 9/7 filter pair to a frame from the "Mother and Daughter" sequence and plotted the logarithm of the signal variance of each subband $\ln \sigma_j^2$ as a function of the scale j. Fig. 3 demonstrates that $\ln \sigma_j^2$ is approximately linearly dependent on the scale number j, i.e.,

$$\ln \sigma_j^2 \cong kj + b. \tag{5}$$

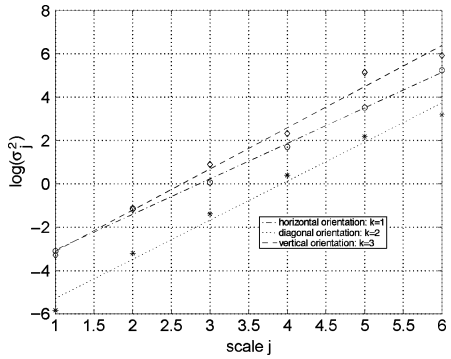


Fig. 3. The logarithm $\ln \sigma_j^2$ as a function of j. The symbols indicate the experimental data and the dotted lines are curves fitted by (5).

It is also seen that the slope k in the last equation is approximately the same for all three orientations. In this case, we found the three slopes to be 1.6358, 1.8002, 1.8969 for the horizontal, diagonal, and vertical orientations by fitting (5) to experimental data. Although the signal variance decays exponentially across successive scales for a natural image or L frame, it should be noted that such a rule does not hold for an H frame. Therefore, the signal variances in different subbands of an H frame should be calculated separately.

B. Bit-Rate Estimation for Embedded Context-Based Entropy Coding of Detail Subbands

Two broad classes of embedded block-coding techniques have been investigated. One is the context-based entropy coding and the other is the embedded extension to the quad-tree coding schemes [27]. Due to the various scalability features of context-based entropy coding, it is adopted by JPEG2000 and most of the latest video coding standard. In embedded quantization followed by intrasubband context-based adaptive entropy coding, each subband is coded independently and represented by an efficient embedded bitstream, where prefixes of the bitstream correspond to successively finer quantization [27]. It is known that uniform scalar quantization is asymptotically optimal for a Laplacian-distributed random variable [17], and that double-deadzone successive-approximation quantization (SAQ) has been the dominant choice for most wavelet video coders. Let Δ be the quantization step size of the bin of a deadzone scalar quantizer with SAQ. The quantization of a sample x is performed as

$$\mathrm{q}(x) = \mathrm{sign}(x) \begin{cases} \left\lfloor \frac{|x|}{\Delta} \right\rfloor, & \frac{|x|}{\Delta} > 0 \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

The last equation demonstrates that all the values in the range $|x| \leqslant \Delta$ are quantized to zero, and hence, the size of the zero bin (deadzone) is twice that of the nonzero bin, i.e., $T_d/\Delta = 1$, where T_d is the threshold of zero bin. This is also proved to be a nearly optimum $(T_d/\Delta \cong 0.81)$ value for wavelet video coders in the sense of minimizing quantization distortion [17]. The ratio of nonzero quantized coefficients is

$$p\left(|X| \geqslant T_d\right) = 1 - \int_{-T_d}^{T_d} p(x)\mathrm{d}x = \mathrm{e}^{-\frac{\sqrt{2}}{\sigma}\Delta} \triangleq \rho. \tag{7}$$

These nonzero quantized coefficients are called the significant coefficients, which are subject to sign and magnitude refinement coding. The distribution of significant coefficients can be derived from $p(x)$ conditioned on nonzero quantization

$$\begin{aligned} p_T(x) &= p\left(x|\mathrm{q}(X) \neq 0\right) = \frac{p\left(x, |X| \geqslant T_d\right)}{p\left(|X| \geqslant T_d\right)} \\ &= \rho^{-1} p(x) \mathbf{I}\left(|x| \geqslant T_d\right) \end{aligned} \tag{8}$$

where $\mathbf{I}(\cdot)$ is the indicator function. When using context-adaptive arithmetic coding, the resulting coding bit rate from the sign and magnitude refinement primitives approaches the conditional entropy of the source $\mathcal{H}(\mathrm{q}(X)|\mathcal{N}X, |X| \geqslant T_d)$, which can be estimated by using (3), (8), and the following proposition.

Proposition 1: The entropy of a quantized significant coefficient conditioned on its neighbors can be estimated as

$$\begin{aligned} &\mathcal{H}\left(\mathrm{q}(X)|\mathcal{N}X, |X| \geqslant T_d\right) \\ &\quad \cong \mathcal{H}\left(\mathrm{q}(X)||X| \geqslant T_d\right) - \mathcal{I}\left(X; \Theta||X| \geqslant T_d\right) \\ &\quad \cong \mathcal{H}\left(\mathrm{q}(X)||X| \geqslant T_d\right) - \frac{0.2988}{(\nu + 0.9773)^{0.8}} \\ &\quad \text{(bits/coefficient)} \end{aligned} \tag{9}$$

where $\nu \triangleq \Delta/\sigma$ and

$$\mathcal{H}\left(\mathrm{q}(X)||X| \geqslant T_d\right) = 1 - \log_2(1/\rho - 1) - (\log_2 \rho)/(1-\rho) \quad \text{(bits/coefficient)} \tag{10}$$

is the entropy of a significant coefficient unconditioned on its neighbors.

Proof: See Appendix I. ■

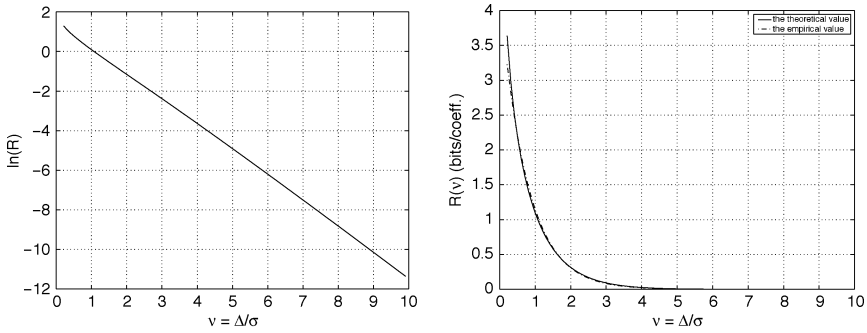
The proof is based on the fact that the mutual information $\mathcal{I}(X; \Theta||X| \geqslant T_d)$ reflects the reduction of bit rate in coding a significant coefficient when knowledge of its neighbors is available.

The mutual information $\mathcal{I}(X; \Theta||X| \geqslant T_d)$ is a convex monotonically decreasing function of ν, which indicates that, as the deadzone threshold increases, the information about a significant wavelet coefficient X from its state variable Θ decreases. This is reasonable, since knowing the state variable Θ barely gives more information when a coefficient X is already known to lie beyond a very large deadzone threshold.

When a wavelet coefficient lies within the quantization deadzone, i.e., $|X| < T_d$, it will be classified as an insignificant coefficient, and its position with respect to the significant coefficients is recorded in a significance map. This is done by the zero coding (ZC) primitive [21], [23]. Notice that, in the majority of cases, the significance map is first run-length coded and then arithmetically coded. Due to the near-entropy performance of arithmetic coding, the coding bit rate from the ZC primitive is very close to the conditional entropy of a binary source, which is 1 when $|X| \geqslant T_d$ and 0 when $|X| < T_d$. This conditional entropy can be estimated as follows.

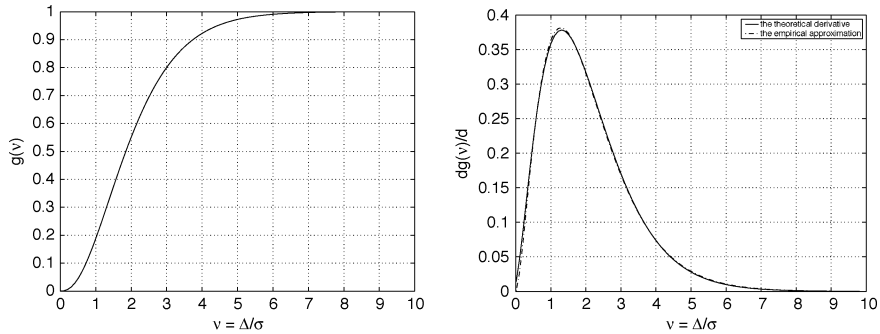
Proposition 2: The bit rate from the ZC coding primitive can be estimated by the formula

$$\mathcal{H}(\mathrm{ZC}|\Theta) = \mathcal{H}(\mathrm{ZC}) - \mathcal{I}(\mathrm{ZC}; \Theta) \tag{11}$$



(a) (b)

Fig. 4. (a) The logarithm $\ln(\mathbf{R}) \sim \nu$, showing a linear relationship. (b) The bit rate $\mathbf{R}$ and its empirical approximation.



(a) $\mathrm{g}(\nu) \sim \nu$ (b) $\mathrm{dg}(\nu)/\mathrm{d}\nu \sim \nu$

Fig. 5. (a) The function $\mathrm{g}(\nu)$. (b) The differentiation of $\mathrm{g}(\nu)$ and its empirical approximation.

where

$$\mathcal{H}(\mathrm{ZC}) = -\rho \log_2 \rho - (1-\rho)\log_2(1-\rho) \quad \text{(bits/coefficient)} \tag{12}$$

and

$$\mathcal{I}(\mathrm{ZC};\Theta) \cong 0.6707\nu \mathrm{e}^{-1.4070\nu} \quad \text{(bits/coefficient)} \tag{13}$$

is the bit-rate saving due to the availability of context information.

Proof: See Appendix II. ■

The total bit rate in coding a subband is therefore

$$\mathbf{R}(\nu) = \rho \mathcal{H}\left(\mathrm{q}(X) | \mathcal{N}X, |X| \geqslant T_d\right) + \mathcal{H}(\mathrm{ZC}|\Theta). \tag{14}$$

The above equation shows that the total bit rate $\mathbf{R}$ is only a function of ν. Thus, we can attempt to derive an empirical approximation to it, similar to the process performed in Appendixes I and II. Fig. 4(a) demonstrates the logarithm of $\mathbf{R}$ as a function of ν. It is seen from this plot that there exists a linear dependence between $\ln \mathbf{R}$ and ν. This suggests we may use a function of the form $k\mathrm{e}^{-\alpha\nu}$ to approximate $\mathbf{R}$, as shown in Fig. 4(b). Fitting the empirical formula to the theoretical expression, we determined the parameters to be $k \cong 4.2469$ and $\alpha \cong 1.3102$, which leads directly to the following proposition:

Proposition 3: The context-based coding bit rate of a wavelet detail subband is approximately

$$\mathbf{R}(\nu) \cong 4.2469\mathrm{e}^{-1.3102\nu} \quad \text{(bits/coefficient)}. \tag{15}$$

C. Quantization Distortion in the Detail Subbands

When the centroid value is chosen for each quantization bin $\widehat{X}$, the expected distortion caused by a uniform quantizer with deadzone $T_d = \Delta$ can be derived as

$$\begin{aligned} \varepsilon &= \mathrm{E}\left[(X - \widehat{X})^2\right] \\ &= \left\{-\rho(\nu + 1/\sqrt{2})^2 + 1 - \rho^2\nu^2/(1-\rho)^2\right\}\sigma^2 \\ &= \mathrm{g}(\nu)\sigma^2 \end{aligned} \tag{16}$$

where

$$\mathrm{g}(\nu) \triangleq -\rho(\nu + 1/\sqrt{2})^2 + 1 - \rho^2\nu^2/(1-\rho)^2$$

is the quantization distortion normalized to the signal variance and is only a function of ν. When the ratio $\nu = \Delta/\sigma$ is fixed, the distortion is proportional to the signal variance σ^2. The normalized distortion $\mathrm{g}(\nu)$ is plotted as a function of ν in Fig. 5(a). It is seen that $\mathrm{g}(\nu)$ approaches one as $\nu \to \infty$, which means that as the quantization step size becomes larger, the distortion approaches the signal variance. The derivative of $\mathrm{g}(\nu)$ is shown in Fig. 5(b), which suggests it may be approximated by the empirical formula $k\nu^{\beta}\mathrm{e}^{-\alpha\nu}$. To simplify the calculation in the next section, we fixed α to be 1.3102 and fit the other two parameters k and β. The fitting results show that $k \cong 1.3474$ $\beta \cong 1.6853$ and the empirical formula is plotted in Fig. 5(b), which indicates a very good approximation. By summarizing the above results, we have the following proposition.

Proposition 4: The distortion of a Laplacian random variable X by a uniform quantizer with deadzone $T_d = \Delta$ is $\varepsilon = \mathrm{g}(\nu)\sigma^2$, where the derivative of $\mathrm{g}(\nu)$ is approximated by the following empirical formula:

$$\mathrm{dg}(\nu)/\mathrm{d}\nu \cong 1.3474\nu^{1.6853}\mathrm{e}^{-1.3102\nu}. \tag{17}$$

Proposition 4 will be used in the calculation of the optimum bit-rate allocation in the next section.

D. Optimum Bit-Rate Allocation and Rate-Distortion Function

Let the subband of the kth ($k = 1, 2, 3$) orientation in scale j be denoted as (j, k) and the coarsest resolution low-frequency subband be J. Due to the orthogonality of the DWT, the average distortion in the original image caused by quantization of its subband coefficients is [27], [36]

$$\begin{aligned}\mathbf{d} &= 4^{-J}G_J\varepsilon_J + \sum_{j=1}^{J}\sum_{k=1}^{3} 4^{-j}G_{j,k}\varepsilon_{j,k} \\ &= \left\{4^{-J}G_J\tilde{\varepsilon}_J + \sum_{j=1}^{J}\sum_{k=1}^{3} 4^{-j}G_{j,k}\tilde{\varepsilon}_{j,k}\right\}\sigma_0^2 \\ &= \tilde{\mathbf{d}}\sigma_0^2 \end{aligned} \tag{18}$$

where $G_{j,k}$ is the synthesis gain associated with subband (j, k) and $\tilde{\mathbf{d}}$ and $\tilde{\varepsilon}$ are the quantization distortions normalized to the average signal variance σ_0^2 of the original frame. Commonly used biorthogonal wavelet filter pairs, such as the 9/7, approximate orthonormality fairly well. Therefore, the synthesis gain is almost unity, i.e., $G_{j,k} \cong 1$. On the other hand, (18) is valid under the assumption that quantization noise is independently manifested within each subband, which is approximate true in most cases. Similarly, the average bit rate is

$$\mathfrak{R} = 4^{-J}\mathbf{R}_J + \sum_{j=1}^{J}\sum_{k=1}^{3} 4^{-j}\mathbf{R}_{j,k}. \tag{19}$$

The objective of real video codec design is to minimize the distortion given a total bit-rate constraint $\mathfrak{R} \leqslant \mathfrak{R}_T$. This is done based on Lagrangian RD optimization, which yields

$$\frac{\partial}{\partial \mathbf{R}_{j,k}}(\tilde{\mathbf{d}} + \lambda\mathfrak{R}) = 0. \tag{20}$$

The convexity of RD function ensures that the minimum exists and the optimum bit rate for every subband is the point where the RD function of all the subbands have equal slope [21], [23]. Substituting (15) and (17)–(19) into (20) yields

$$\widehat{\nu}_{j,k}(\tilde{\sigma}_{j,k}, \lambda) \cong \left[\frac{4.1296\lambda}{\tilde{\sigma}_{j,k}^2}\right]^{0.5934} \tag{21}$$

where $\tilde{\sigma}_{j,k}^2$ is the signal variance of the corresponding subband (j, k), normalized to σ_0^2. The last equation gives the optimum value of the quantization step-size signal variance ratio ν for detail subband (j, k) under the approximations mentioned before. This is also valid for low-frequency subbands of the H frames, since the model of (1) holds for them as well, as shown in Section II-A.

To estimate the optimum value of ν in an L frame, we assume the quantizer works in low distortion region. Shannon's upper bound can be used to approximate the RD function of this subband [27]

$$\begin{aligned}\mathbf{R}_J &\cong \log_2\sqrt{2\pi\mathrm{e}} - \log_2\nu \\ \varepsilon_J &\cong \frac{\nu^2}{12}\sigma_J^2.\end{aligned}$$

The optimum value of ν is therefore

$$\widehat{\nu}_J(\tilde{\sigma}_J, \lambda) \cong \sqrt{\frac{6\lambda}{\tilde{\sigma}_J^2 \ln 2}} = \sqrt{\frac{8.6562\lambda}{\tilde{\sigma}_J^2}} \tag{22}$$

where $\tilde{\sigma}_J$ is the variance of temporal low-frequency subband J normalized to σ_0^2.

The Lagrange parameter λ is strictly positive and controls both the bit rate and distortion of each subband. It is seen that as the subband variance increases, the quantization step size must decrease in order to reduce the distortion, which is consistent with the standard codec design. Finally, the quantization distortion of a coded frame and its coding bit rate are related by the parameter λ, i.e., for each $\lambda > 0$, the distortion $\mathbf{d}$ and the total bit rate $\mathfrak{R}$ are related by the value $\widehat{\nu}(\tilde{\sigma}, \lambda)$ through (18) and (19). Thus, the pair $\{\mathbf{d}(\lambda), \mathfrak{R}(\lambda)\}$ provides an approximate description of the operational RD curve of a wavelet codec in coding a single frame.

III. Propagation of Quantization Noise in the Temporal Decomposition Tree

In this section, we analyze the average frame distortion at different temporal levels for a t + 2D coding scheme. The RD behavior of a 2D + t structure can be derived by a similar analysis as the one outlined below.

Current wavelet video coders typically use the Haar and 5/3 filter pairs for MCTF. Other filters, such as the longer 9/7 filter pair, can also be used to exploit the dependency between successive video frames and hence improve the RD performance. We will analyze the simplest Haar filter first and then generalize the results to more complicated temporal filters.

A. Distortion Distribution for Haar Temporal Filtering

In a temporal filtering structure with T levels, there are 2^T frames in one group of frames (GOF). Assume that, within a GOF, the signal variance E_0 is approximately constant and let A and B stand for the even and odd frames in the motion-compensated lifting structure [15]. The temporal high-frequency frame H coincides with frame A, and the temporal low-frequency frame L coincides with frame B (refer to Fig. 6).

During motion compensation, the pixels can be classified into three types: connected, unconnected, and multiple connected. For most video sequences, frame A has only a small fraction of pixels that do not have a correspondence in the reference frame and are declared as intracoded pixels [30]. The intracoded pixels must be treated differently due to their nondifferential signal statistics. This is problematic in 3-D wavelet

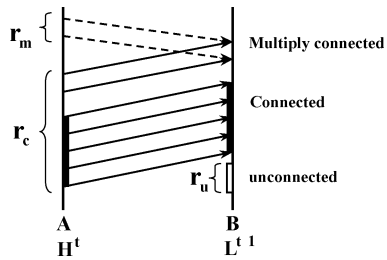


Fig. 6. Motion estimation for Haar filters. The pixels can be classified into three types: multiple connected, connected, and unconnected.

coding since the 2-D DWT used for the spatial decomposition is a global transform. Several solutions exist for intraprediction of nondifferential areas in the error frames based on the causal (i.e., already processed) neighborhood around them. However, the adoption of such a strategy would specialize our results to a certain realization. As a result, we opted to omit such an analysis in our current derivations and assume that all pixels in the A frame are residues from motion-compensated prediction. Correspondingly, all the pixels in frame A fall into two categories: connected and multiple connected. For connected pixels, there exists a motion vector (k_c, l_c) that maps motion trajectory from frame A to B with the inverse motion vector $(\tilde{k}_c, \tilde{l}_c) = (-k_c, -l_c)$ mapping motion trajectory back from frame B to frame A. The connections corresponding to the multiple connected pixels in frame A are only applied for temporal prediction and are not considered during low-pass filtering. For this reason, the multiple connected pixels in frame A compose only predicted areas. The multiple connected pixels in frame B have one unique connection to frame A employing only one estimated motion trajectory, and therefore are treated like connected pixels [30], [35]. Let r_c be the ratio of connected pixels, r_m be the multiple connected pixels, and r_u be the ratio of unconnected pixels. Due to the above analysis, we have $r_c + r_m = 1$ and $r_u = r_m$, as shown in Fig. 6.

The propagation of quantization noises along the wavelet decomposition tree has been intensively studied by Rusert *et al.* [30], [36]. For completeness, we review their results prior to deriving the average distortion in different temporal levels. In the following, superscript (t) denotes the tth temporal decomposition level. Following the method of [30] and based on our previous assumptions for connected and unconnected pixels, the average distortion of the reconstructed pair of frames A and B can be expressed as

$$\mathbf{d}_L^{(0)} = \frac{1}{2}(\mathbf{d}_A + \mathbf{d}_B) = \left(\frac{3}{4} - \frac{r_c}{4}\right)\mathbf{d}_H^{(1)} + \frac{1}{2}\mathbf{d}_L^{(1)} \tag{23}$$

where $\mathbf{d}_L^{(t)}$ and $\mathbf{d}_H^{(t)}$ denote the quantization distortion in the reconstructed frame $L^{(t)}$ and $H^{(t)}$, i.e., the temporal low- and high-frequency frames in the tth temporal level, and the values of $\mathbf{d}_L^{(t)}$ and $\mathbf{d}_H^{(t)}$ are given by (18). The derivation of (23) is given in Appendix III. By averaging both sides of (23), we obtain the relationship between the average frame distortion at temporal levels 0 and 1

$$\bar{\mathbf{d}}_L^{(0)} = \left(\frac{3}{4} - \frac{\bar{r}_c}{4}\right)\mathbf{d}_H^{(1)} + \frac{1}{2}\bar{\mathbf{d}}_L^{(1)} \tag{24}$$

where the bar indicates the average value. Generally, the average distortion of L frames in the tth temporal level is given by

$$\bar{\mathbf{d}}_L^{(t-1)} = \left(\frac{3}{4} - \frac{\bar{r}_c(t)}{4}\right)\mathbf{d}_H^{(t)} + \frac{1}{2}\bar{\mathbf{d}}_L^{(t)} \tag{25}$$

which leads to

$$\bar{\mathbf{d}}_L^{(0)} = \sum_{t=1}^{T}\left(\frac{3}{4} - \frac{\bar{r}_c(t)}{4}\right)\left(\frac{1}{2}\right)^{t-1}\mathbf{d}_H^{(t)} + \left(\frac{1}{2}\right)^{T}\bar{\mathbf{d}}_L^{(T)} \tag{26}$$

where $\mathbf{d}_L^{(t)}$ and $\mathbf{d}_H^{(t)}$ are calculated by (18).

The average bit rate for one GOF is calculated by averaging the frame bit rates

$$\overline{\mathfrak{R}} = 2^{-T}\mathfrak{R}_L + \sum_{t=1}^{T} 2^{-t}\mathfrak{R}_H^{(t)} + \mathfrak{R}_{mv} \tag{27}$$

where $\mathfrak{R}_L$ and $\mathfrak{R}_H^{(t)}$ represent the coding bit rate for the L and H frames in temporal level t and are evaluated by (19); $\mathfrak{R}_{mv}$ is the bit rate for motion vectors. It is important to notice that $\mathfrak{R}_{mv}$ can significantly affect the RD performance, as well as the resulting complexity [29]. In the case that no motion vector scalability is used, $\mathfrak{R}_{mv}$ is a fixed value. Equations (26) and (27) approximately describe the operational RD behavior for the Haar filter pair case.

B. Generalization to Longer Temporal Filters

The derivation given in the previous section is only valid under the assumption of accurate invertibility of the motion trajectories. However, this assumption does not hold for cases such as subpixel interpolation [23], [35], [36]. On the other hand, motion-compensated lifting structures for longer filters such as the 5/3 and 9/7 filter pairs are much more complicated than that for the Haar filter pair, which makes it difficult to track the quantization noise along the MCTF tree. Nevertheless, (25) suggests that we can always find a linear relationship between the average frame distortions within adjacent temporal levels

$$\bar{\mathbf{d}}_L^{(t)} = A^{(t+1)}\bar{\mathbf{d}}_L^{(t+1)} + B^{(t+1)}\mathbf{d}_H^{(t+1)}. \tag{28}$$

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

$$\begin{aligned}\bar{\mathbf{d}}_L^{(0)} &= \sum_{t=1}^{T} B^{(t)}\prod_{j=1}^{t-1} A^{(j)}\mathbf{d}_H^{(t)} + \prod_{t=1}^{T} A^{(t)}\bar{\mathbf{d}}_L^{(T)} \\ &= [\mathcal{A}\mathcal{B}_1, \ldots, \mathcal{B}_T]\left[\bar{\mathbf{d}}_L^{(T)}\mathbf{d}_H^{(1)}, \ldots, \mathbf{d}_H^{(T)}\right]^T\end{aligned} \tag{29}$$

where $\mathcal{A} \triangleq \prod_{t=1}^{T} A^{(t)}$, $\mathcal{B}_t \triangleq B^{(t)} \prod_{j=1}^{t-1} A^{(j)}$, $t = 1, \ldots, T$, are the linear coefficients, which depend on the motion estimation algorithm, the wavelet filter pair used for MCTF, and the video sequence characteristics. Notice that (26) is a special form of (29), with $B^{(t)} = ((3/4) - (\bar{r}_c(t)/4),)$, $A^{(t)} = 1/2$, since the Haar filter pair is the simplest instantiation of MCTF.

For long temporal filters, the number of L and H frames remains the same within one GOF as in the Haar case; hence, the average bit rate in this case is given by (27).

IV. Modeling Procedure

The average distortion in the reconstructed video sequence is a linear combination of the distortion in the H and L frames, as shown in (29). Consequently, it is also a linear combination of the distortion in all different subbands. On the other hand, the distortion is closely related to the subband variances, which can be calculated very efficiently because of their regular distribution within L and H frames. Given the Lagrange parameter λ, both the average coding bit rate and the distortion can be estimated by (27) and (29). Therefore, the modeling procedure can be summarized as follows.

- Calculate each subband variance for L and H frames.
  1) The subband signal variance σ_j^2 decays approximately exponentially across scales in an L frame. Using this property, we first choose two detail subbands for each of the three orientations, calculate the corresponding signal variance, and then estimate the signal variance in other detail subbands by (5). The signal variance in the coarsest representation subband is calculated separately for each L frame. After calculating all these values $\sigma_{j,k}^2$, normalize them to the signal variance σ_0^2 in the original L frame to get $\tilde{\sigma}_{j,k}^2$.
  2) To calculate the normalized signal variance distribution $\tilde{\sigma}_{j,k}^2$ of H frames in temporal level k, we choose several H frames in this temporal level, calculate the normalized signal variance for each subband of the selected frames, and then average the distribution of these frames for $\tilde{\sigma}_{j,k}^2$.
- Calculate the bit rate of L and H frames.
  1) Use the signal variance distribution $\tilde{\sigma}_{j,k}^2$ to calculate the optimum value of $\widehat{\nu}_{j,k}$ by (21) and (22).
  2) Calculate the coding bit rate of a frame for $\lambda > 0$ using (15) and (19).
  3) Calculate the average frame bit rate using (27).
- Evaluate the average distortion.
  1) Use the optimum value $\widehat{\nu}_{j,k}$ to evaluate the frame distortion for $\lambda > 0$ using (16) and (18).
  2) Adapt the linear coefficients in (29) to fit the RD curve to the offline training data.

V. Experimental Results

By following the above procedure, we optimized the parameters in the model with respect to our experimental data for three video sequences at Common Immediate Format (CIF) resolution: "Coastguard," "Akiyo," "Mother and Daughter." Each sequence is compressed at a frame rate of 30 Hz in two scenarios by Microsoft SVC [32], which is an improved version of the codec described in [23]. The experimental RD data are obtained by averaging the frame distortion over 256 frames for different bit rates.

In the first scenario, the temporal decomposition level is set to four and the spatial decomposition level is varied from one to three. The Daubechies 5/3 filters are used for both temporal and spatial decomposition. The codec is set to t + 2D MCTF mode and the frame rate is set to 30 Hz. The bitstream is truncated at the following values: 128, 384, 512, 768, 1024, 1280, 1536 (kbps). The distortion is expressed in terms of peak SNR (PSNR). The model is first trained to experimental data, and once this is done, the model can be used to make predictions of real operation RD performance for various coding parameters. In this experiment, 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 are shown with symbols. The results in Fig. 7 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 fewer bits are needed to achieve a better PSNR. But at higher bit rates, a higher spatial decomposition level is not always a better choice.

In the other scenario shown by Fig. 8, we fixed the spatial decomposition level to three while changing the temporal filtering level from two to four. Again the model is adapted using three training data points, and it successfully predicts the other experimental data points.

VI. Conclusion

We developed an analytical RD model for a t + 2D wavelet video codec. The subbands are coded independently and the bit rates of different subbands are optimally truncated to minimize the distortion under a bit-rate constrained scalable coding. Due to the remaining intrascale dependency of wavelet coefficients, the subband bit rate can be greatly reduced by using context-based coding. The bit-rate savings can be estimated from the doubly stochastic model, which accurately captures the dependency of wavelet coefficients both for L and H frames. Our model demonstrates that the average distortion that stems from coding a video sequence is the linear combination of the distortion in coding the subbands. By adapting the coefficients in this linear model to offline training data, we can predict the operational RD performance at encoding time very accurately.

Appendix I
Proof of Proposition 1

We first derive (10). According to (8), the probability of a significant coefficient's falling into the ith quantization bin is

$$p_i = \int_{T_d+(i-1)\Delta}^{T_d+i\Delta} p_T(x)\mathrm{d}x = \frac{1}{2}\left(\frac{1}{\rho} - 1\right)\rho^i, \quad i \neq -1, 0.$$

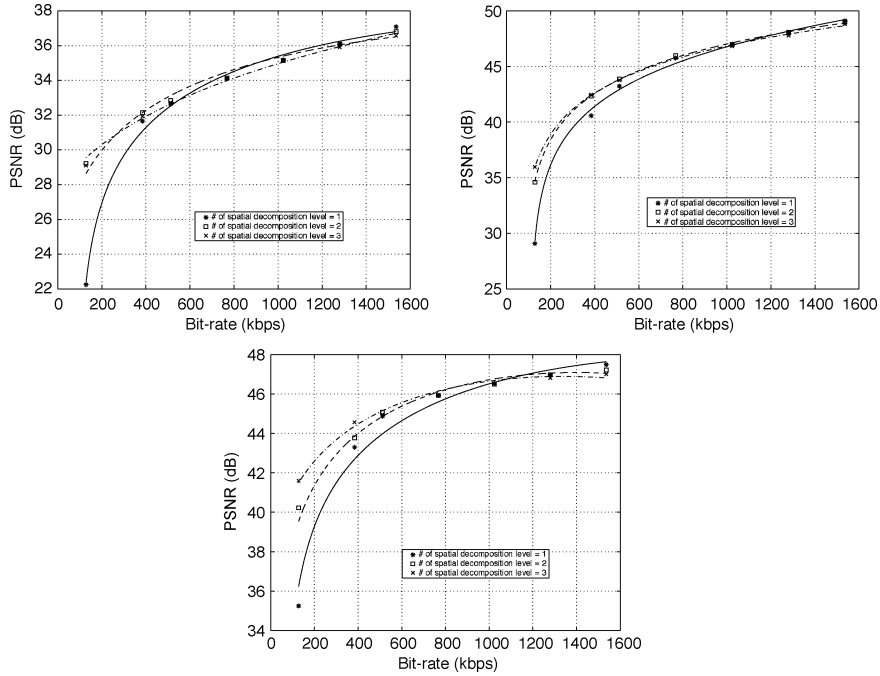


Fig. 7. Operational RD curves of three video sequences. The spatial decomposition level varies from one to three with the temporal decomposition levels fixed to be four. The symbols indicate the experimental data in different scenarios and the curves are fitted by the theoretical model. The video sequences from top to bottom: "Coastguard," "Akiyo," and "Mother and Daughter."

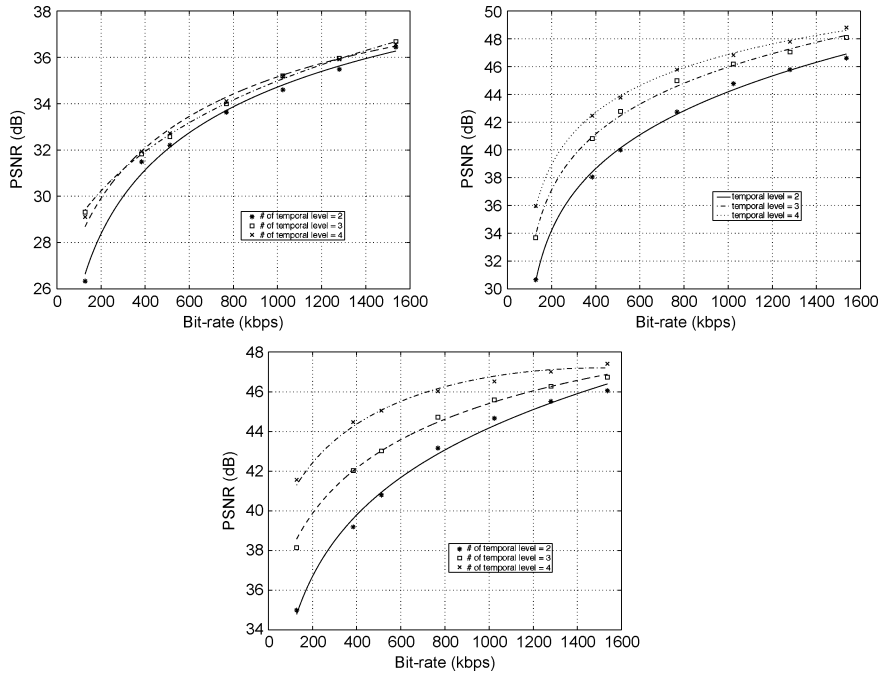


Fig. 8. Operational RD curves of three video sequences. The temporal decomposition level varies from two to four and the spatial decomposition level is fixed to be three. The symbols indicate the experimental data in different scenarios and the curves are fitted by the theoretical model. The video sequences from top to bottom: "Coastguard," "Akiyo," and "Mother and Daughter."

From the definition of entropy, we have

$$\begin{aligned}&\mathcal{H}\left(\mathrm{q}(X)\middle||X|\geqslant T_d\right)\\&\quad=-\sum_{i\neq-1,0}p_i\log_2 p_i\\&\quad=1-\log_2(1/\rho-1)-(\log_2\rho)/(1-\rho)\ \text{(bits/coefficient)}\end{aligned}$$

which is (10). It should be noted that the last equation is the entropy of a quantized significant coefficient unconditioned on its neighbors. Since most video codecs adopt context-based encoding, where each bit plane is coded adaptively on the observation of existing bit planes, (10) is an overestimate of the coding bit rates from the sign and magnitude coding primitives. It is known that in the low-distortion region, the bit-rate reduc-

tion in coding a random variable X obtained by the knowledge of another random variable Y is their mutual information $\mathcal{I}(X;Y)$ [25], so we may estimate the coding bit rates from a context-adaptive encoder by the following equation:

$$\mathcal{H}(\mathrm{q}(X)|\mathcal{N}X, |X| \geqslant T_d) \cong \mathcal{H}(\mathrm{q}(X)||X| \geqslant T_d) - \mathcal{I}(X;\mathcal{N}X||X| \geqslant T_d).$$

Notice that, based on (3) $X \longrightarrow \Theta \longrightarrow \mathcal{N}X$ forms a Markov chain; therefore

$$\mathcal{I}(X;\Theta||X| \geqslant T_d) \geqslant \mathcal{I}(X;\mathcal{N}X||X| \geqslant T_d) \tag{30}$$

according to the property of Markov chains. When enough neighboring coefficients are taken into account, the state Θ of the current coefficient is known from its maximum a posteriori (MAP) estimation $\widehat{\Theta}_{\mathrm{MAP}}$. The state variables Θ are typically very closely related within adjacent neighbors [18]. Therefore, (30) will assume equality under the condition that sufficient statistics from $\mathcal{N}X$ are available for estimating Θ

$$\begin{aligned}\mathcal{I}(X;\mathcal{N}X||X| \geqslant T_d) &= \mathcal{I}\left(X;\widehat{\Theta}_{\mathrm{MAP}}||X| \geqslant T_d\right) \\ &\cong \mathcal{I}(X;\Theta||X| \geqslant T_d).\end{aligned} \tag{31}$$

It is known that, under the assumption of doubly stochastic model (1), the sufficient statistics from $\mathcal{N}X$ for estimating Θ do exist and take the form [25]

$$\mathcal{T}(\mathcal{N}X) = \sum_{i \in \mathcal{N}X} w_i \mathcal{N}X_i^2 \tag{32}$$

where w_i is the weight for $\mathcal{N}X_i$. On the other hand, in most video codecs, a large number of neighbor coefficients are used in context-adaptive coding. For example, 17 contexts are used in [23] for sign and magnitude refinement primitives, while eight contexts are used in [21]. Therefore, Θ can be estimated accurately and the above conditions for equality are approximately met.

Notice that (31) enables us to estimate the bit-rate reduction due to the dependency among wavelet coefficients. According to the definition of mutual information

$$\mathcal{I}(X;\Theta||X| \geqslant T_d) = \mathcal{H}(X||X| \geqslant T_d) - \mathcal{H}(X|\Theta, |X| \geqslant T_d). \tag{33}$$

Next we will derive the two parts on the right-hand side of (33) separately. To simplify the derivation, the random variables X and Θ are linearly mapped to X' and Θ' by the following transform:

$$\begin{bmatrix} X' \\ \Theta' \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma} & 0 \\ 0 & \frac{1}{\sigma^2} \end{bmatrix} \begin{bmatrix} X \\ \Theta \end{bmatrix}. \tag{34}$$

The mutual information will be invariant under any invertible linear transform, and hence

$$\begin{aligned}&\mathcal{I}(X;\Theta||X| \geqslant T_d) \\ &\quad = \mathcal{I}(X';\Theta'||X| \geqslant T_d) \\ &\quad = \mathcal{H}(X'||X| \geqslant T_d) - \mathcal{H}(X'|\Theta', |X| \geqslant T_d).\end{aligned} \tag{35}$$

We first derive the following conditional probabilities that will appear in the calculation of (31). Using (8), we have the probability density of X' conditioned on nonzero quantization

$$p(x'||X| \geqslant T_d) = \sigma p(x||X| \geqslant T_d) = \frac{1}{\sqrt{2}\rho} \mathrm{e}^{\sqrt{2}|x'|} \mathbf{I}(|x'| \geqslant \nu) \tag{36}$$

where $\nu \triangleq \Delta/\sigma$. Similarly, the probability density of X' conditioned on both the state Θ' and the nonzero quantization is

$$p(x'|\theta', |X| \geqslant T_d) = \frac{p(x'|\theta')\mathbf{I}(|x'| \geqslant \nu)}{2\mathrm{Q}\left(\frac{\nu}{\sqrt{\theta'}}\right)} \tag{37}$$

where $p(x'|\theta') = (1/\sqrt{2\pi\theta'})/\mathrm{e}^{-(x'^2/2\theta')}$ and $\mathrm{Q}(t) \triangleq (1/\sqrt{2\pi}) \int_t^\infty \mathrm{e}^{-(x^2/2)} \mathrm{d}x$.

Finally, the probability density of Θ' conditioned on nonzero quantization is

$$p(\theta'||X| \geqslant T_d) = \frac{p(|X| \geqslant T_d|\theta')\, p(\theta')}{p(|X| \geqslant T_d)} = \rho^{-1} 2\mathrm{Q}\left(\frac{\nu}{\sqrt{\theta'}}\right) p(\theta'). \tag{38}$$

Based on (36), the first part of (35) becomes

$$\begin{aligned}&\mathcal{H}(X'||X| \geqslant T_d) \\ &\quad = -\int p(x'||X| \geqslant T_d) \ln\left[p(x'||X| \geqslant T_d)\right] \mathrm{d}x' \\ &\quad = -\int_{|x'| \geqslant \nu} \frac{1}{\sqrt{2}\rho} \mathrm{e}^{-\sqrt{2}|x'|} \ln\left[\frac{1}{\sqrt{2}\rho} \mathrm{e}^{-\sqrt{2}|x'|}\right] \mathrm{d}x' \\ &\quad = \ln(\sqrt{2}\mathrm{e}) \quad (\mathrm{nats}).\end{aligned} \tag{39}$$

The second part is

$$\begin{aligned}&\mathcal{H}(X'|\Theta', |X| \geqslant T_d) \\ &\quad = -\int\int p(\theta'||X| \geqslant T_d)\, p(x'|\theta', |X| \geqslant T_d) \\ &\qquad \times \ln p(x'|\theta', |X| \geqslant T_d) \mathrm{d}x' \mathrm{d}\theta' \\ &\quad = \rho^{-1} \left\{ \frac{\nu}{\sqrt{2}\exp\{\sqrt{2}\nu\}} + \ln(2\pi\mathrm{e}) \int_0^\infty \mathrm{e}^{-\theta'} \mathrm{Q}\left(\frac{\nu}{\sqrt{\theta'}}\right) \mathrm{d}\theta' \right. \\ &\qquad + \int_0^\infty \mathrm{e}^{-\theta'} \mathrm{Q}\left(\frac{\nu}{\sqrt{\theta'}}\right) \ln\theta' \mathrm{d}\theta' \\ &\qquad \left. + \int_0^\infty \mathrm{e}^{-\theta'} 2\mathrm{Q}\left(\frac{\nu}{\sqrt{\theta'}}\right) \ln\left[2\mathrm{Q}\left(\frac{\nu}{\sqrt{\theta'}}\right)\right] \mathrm{d}\theta' \right\} \\ &\qquad (\mathrm{nats}).\end{aligned} \tag{40}$$

$\mathcal{I}(X;\Theta||X| \geqslant T_d)$ is only a function of ν since $\mathcal{H}(X'||X| \geqslant T_d)$ is a constant and $\mathcal{H}(X'|\Theta', |X| \geqslant T_d)$ is a function of ν in (40). But the difficulty lies in the fact that there exists no explicit form for the integration in (40). Results from numerical integration show that $\ln[\mathcal{I}(X;\Theta||X| \geqslant T_d)]$ decays at a speed comparable to $-\ln v$, which suggests that $\mathcal{I}(X;\Theta||X| \geqslant T_d)$ may be approximated by an empirical function taking the form $k/(\nu+\beta)^{\alpha}$. By fitting this empirical function to the numerical calculation results, we got the value of the parameters $k \cong 0.2071$, $\beta \cong 0.9773$, $\alpha \cong 0.8098$. Therefore

$$\mathcal{I}(X;\Theta||X| \geqslant T_d) \cong \frac{0.2071}{(\nu+0.9773)^{0.8}} \quad (\text{nats}). \tag{41}$$

Fig. 9 shows the mutual information $\mathcal{I}(X;\Theta||X| \geqslant T_d)$ together with its empirical approximation on the same plot. The empirical function matches the real entropy very closely, indicating that (41) is a very accurate approximation. Inserting (39) and (41) into (35) and changing the unit to bits/coefficient yields (9).

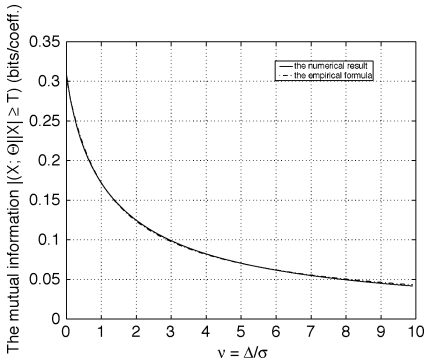


Fig. 9. The mutual information $\mathcal{I}(X;\Theta||X| \geqslant T_d)$ and its empirical approximation.

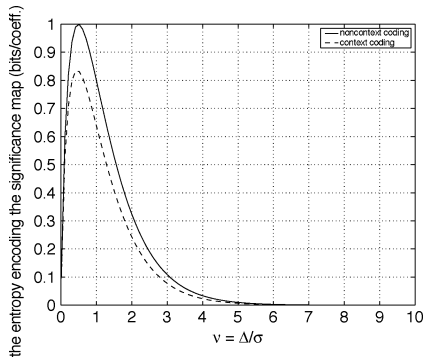


Fig. 10. The entropy of zero-coding primitive with and without taking context into account.

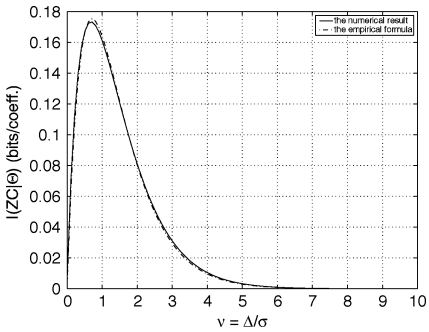


Fig. 11. The reduction of bit rate due to context zero-coding. The figure shows both the numerical calculation and its empirical approximation.

Appendix II
Proof of Proposition 2

Without considering the redundancy among the wavelet coefficients, the output entropy of the ZC primitive will be

$$\mathcal{H}(\text{ZC}) = -\rho\ln(\rho) - (1-\rho)\ln(1-\rho) \quad (\text{nats}). \tag{42}$$

In real video codecs, the redundancy among the wavelet coefficients is used to reduce the bit rate in coding the significance map, and the reduction is the mutual information obtained from the neighbor of the current coefficient $\mathcal{I}(\text{ZC}, \mathcal{N}X)$. Following the same steps as in Appendix I, we have the following approximation:

$$\mathcal{I}(\text{ZC};\mathcal{N}X) \cong \mathcal{I}(\text{ZC};\Theta) = \mathcal{H}(\text{ZC}) - \mathcal{H}(\text{ZC}|\Theta). \tag{43}$$

The second term in the above equation can be calculated by using the conditional probability density function $p(x|\theta)$ given by (1)

$$\begin{aligned}\mathcal{H}(\text{ZC}|\Theta) &= -\int_0^{\infty} \frac{1}{\sigma^2}\mathrm{e}^{-\frac{\theta}{\sigma^2}} \left[p\left(|X| \geqslant T_d|\theta\right)\ln p\left(|X| \geqslant T_d|\theta\right)\right. \\ &\quad \left. + \left(1-p\left(|X| \geqslant T_d|\theta\right)\right)\ln\left[\left(1-p\left(|X| \geqslant T_d|\theta\right)\right)\right]\right]\mathrm{d}\theta \\ &= -\int_0^{\infty} \mathrm{e}^{-\theta'}\left[2\mathrm{Q}\left(\frac{\nu}{\sqrt{\theta'}}\right)\ln\left[2\mathrm{Q}\left(\frac{\nu}{\sqrt{\theta'}}\right)\right]\right. \\ &\quad + \left(1-2\mathrm{Q}\left(\frac{\nu}{\sqrt{\theta'}}\right)\right) \\ &\quad \left.\times\ln\left[1-2\mathrm{Q}\left(\frac{\nu}{\sqrt{\theta'}}\right)\right]\right]\mathrm{d}\theta' \quad (\text{nats}).\end{aligned} \tag{44}$$

Again there is no explicit form for the integration in (44). But it should be noted that both $\mathcal{H}(\text{ZC})$ and $\mathcal{H}(\text{ZC}|\Theta)$ (and hence, $\mathcal{I}(\text{ZC};\Theta)$) are only functions of ν, which means we can find an empirical approximation to $\mathcal{I}(\text{ZC};\Theta)$. In Fig. 10, the entropy of ZC primitive is plotted for the cases of both context-based and non-context-based encoding, and knowledge of context information improves the coding efficiency. Fig. 11 plots the difference between the two curves in Fig. 10, i.e., $\mathcal{I}(\text{ZC};\mathcal{N}X)$, indicating the bonus of bit saving from the redundancy of wavelet coefficients. When viewed in the logarithmic scale, the function $\ln[\mathcal{I}(\text{ZC};\mathcal{N}X)]$ shows a decay rate comparable to that of $\ln\nu - \alpha\nu$, indicating it can be approximated by an empirical form $k\nu\mathrm{e}^{-\alpha\nu}$. Fitting this empirical form yields $k \cong 0.6707$ and $\alpha \cong 1.4070$, and hence

$$\mathcal{I}(\text{ZC};\Theta) \cong 0.6707\nu\mathrm{e}^{-1.4070\nu} \quad (\text{bits/coefficient}).$$

APPENDIX III
DERIVATION OF THE AVERAGE DISTORTION FOR A AND B FRAMES

Let $e_{L^{(t)}}$ and $e_{H^{(t)}}$ be the quantization noises in the reconstructed L and H frames in the tth temporal level. Then the quantization noise of the reconstructed connected pixels in the B frame is

$$e_{B_c} = \frac{1}{\sqrt{2}}\left[e_{L^{(1)}} - e_{H^{(1)}}\right]$$

while the quantization noise of the unconnected pixels is

$$e_{B_u} = \frac{1}{\sqrt{2}} e_{L^{(1)}}.$$

In the above expressions, the factor of $\sqrt{2}$ comes from the analysis and synthesis functions of the Haar filter pair. The quantization distortion for these two kinds of pixels is therefore

$$\begin{aligned} |e_{B_c}|^2 &= \frac{1}{2}\left(|e_{L^{(1)}}|^2 + |e_{H^{(1)}}|^2\right) \\ &= \frac{1}{2}\left(\mathbf{d}_L^{(1)} + \mathbf{d}_H^{(1)}\right) \qquad (45a) \\ |e_{B_u}|^2 &= \frac{1}{2}\mathbf{d}_L^{(1)}. \qquad (45b) \end{aligned}$$

From (45), we have the average distortion in frame B

$$\begin{aligned} \mathbf{d}_B &= r_u \frac{1}{2}\mathbf{d}_L^{(1)} + (1-r_u)\frac{1}{2}\left(\mathbf{d}_L^{(1)} + \mathbf{d}_H^{(1)}\right) \\ &= \frac{1}{2}(1-r_u)\mathbf{d}_H^{(1)} + \frac{1}{2}\mathbf{d}_L^{(1)}. \qquad (46) \end{aligned}$$

Similarly, the quantization noise of the connected pixels in the reconstructed frame A is

$$e_{A_c} = \frac{1}{\sqrt{2}}\left[e_{L^{(1)}} + e_{H^{(1)}}\right].$$

As for the multiple connected pixels, the worst case is when the reconstructed frame B is used to estimate the frame A, which introduces an error of $(1/\sqrt{2})[e'_{L^{(1)}} - e'_{H^{(1)}}]$ [36]. This component together with the original quantization error gives

$$e_{A_m} = \frac{1}{\sqrt{2}}\left[e'_{L^{(1)}} - e'_{H^{(1)}}\right] + e_{H^{(1)}}.$$

Therefore, the average distortion in the frame A is

$$\begin{aligned} \mathbf{d}_A &= r_c \frac{1}{2}\left(\mathbf{d}_L^{(1)} + \mathbf{d}_H^{(1)}\right) + (1-r_c)\left(\frac{1}{2}\mathbf{d}_L^{(1)} + \frac{3}{2}\mathbf{d}_H^{(1)}\right) \\ &= \frac{1}{2}\mathbf{d}_L^{(1)} + \left(\frac{3}{2} - r_c\right)\mathbf{d}_H^{(1)}. \qquad (47) \end{aligned}$$

By taking the average of (46) and (47), we get the average distortion of the pair of A and B frames

$$\mathbf{d}_L^{(0)} = \frac{1}{2}\left(\mathbf{d}_A + \mathbf{d}_B\right) = \left(\frac{3}{4} - \frac{r_c}{4}\right)\mathbf{d}_H^{(1)} + \frac{1}{2}\mathbf{d}_L^{(1)}.$$

This is exactly (23). Here $\mathbf{d}_L^{(0)}$ represents the distortion of L frames in the zeroth temporal level, i.e., at the original sequence (A and B frames).

ACKNOWLEDGMENT

The authors would like to thank Dr. Y. Andreopoulos for his helpful suggestion. They also would like to thank the anonymous reviewers of this manuscript for their constructive comments.

REFERENCES

[1] A. M. Gerrish and P. M. Schultheiss, "Information rates of non-Gaussian processes," *IEEE Trans. Inf. Theory*, vol. IT-10, pp. 265–271, Oct. 1964.
[2] D. J. Sakrison, "A geometric treatment of the source encoding of a Gaussian random variable," *IEEE Trans. Inf. Theory*, vol. IT-14, pp. 481–486, May 1968.
[3] ——, "The rate distortion function of a Gaussian process with a weighted square error criterion," *IEEE Trans. Inf. Theory*, vol. IT-14, pp. 506–508, May 1968.
[4] T. Berger, "Information rates of Wiener processes," *IEEE Trans. Inf. Theory*, vol. IT-16, pp. 265–271, Mar. 1970.
[5] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE J. Sel. Areas Commun.*, vol. SAC-5, pp. 1140–1154, Aug. 1987.
[6] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 674–693, Jul. 1989.
[7] 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 Process. Image Commun.*, vol. 4, pp. 477–488, 1992.
[8] B. Girod, "Motion-compensating prediction with fractional-pel accuracy," *IEEE Trans. Commun.*, vol. 41, pp. 604–612, Apr. 1993.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] J. B. O Neal, Jr. and T. Raj Natarajan, "Coding isotropic images," *IEEE Trans. Inf. Theory*, vol. 23, pp. 697–707, Nov. 1997.
[14] A. Hjørungnes and J. M. Lervik, "Jointly optimal classification and uniform threshold quantization in entropy constrained subband image coding," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, Apr. 1997, pp. 3109–3112.
[15] I. Daubechies and W. Sweldens, "Factorization wavelet transforms into lifting steps," *J. Fourier Anal. Appl.*, vol. 4, pp. 247–269, 1998.
[16] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.
[17] S. Mallat and F. Falzon, "Analysis of low bit rate image transform coding," *IEEE Trans Signal Process.*, vol. 46, no. 4, pp. 1027–1042, Apr. 1998.
[18] M. K. Mihçcak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Process. Lett.*, vol. 6, no. 12, pp. 300–303, Dec. 1999.
[19] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Trans. Image Process.*, vol. 9, pp. 1661–1666, Oct. 2000.
[20] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE J. Sel. Areas Commun.*, vol. 18, pp. 1012–1032, Jun. 2000.
[21] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Trans. Image Process.*, vol. 9, pp. 1158–1170, July 2000.
[22] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Appl. Comput. Harmonic Anal.*, vol. 11, pp. 89–123, 2001.
[23] J. Xu, Z. Xiong, S. Li, and Y. Zhang, "Three-dimensional embedded subband coding with optimized truncation (3-D ESCOT)," *Appl. Commun. Harmonic Anal.*, vol. 10, pp. 290–315, 2001.

[24] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden Markov models," *IEEE Trans. Image Process.*, vol. 10, pp. 1056–1068, Jul. 2001.
[25] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Process.*, vol. 10, pp. 1647–1658, Nov. 2001.
[26] 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.
[27] D. S. Taubman and M. W. Marcellin, *JPEG 2000-Image Compression Fundamentals, Standards and Practice*. Norwell, MA: Kluwer Academic, 2002.
[28] Z. He and S. K. Mitra, "A linear source model and a unified rate control algorithm for DCT video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, pp. 970–982, Nov. 2002.
[29] D. Turaga, M. van der Schaar, and B. Pesquet, "Temporal prediction and differential coding of motion vectors in the MCTF framework," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2003.
[30] T. Rusert, K. Hanke, and J. Ohm, "Transition filtering and optimized quantization in interframe wavelet video coding," in *Proc. SPIE Visual Communications and Image Processing (VCIP)*, 2003, vol. 5150, pp. 682–693.
[31] A. T. Deever and S. S. Hemami, "Efficient sign coding and estimation of zero-quantized coefficients in embedded wavelet image codecs," *IEEE Trans. Image Process.*, vol. 12, pp. 420–430, Apr. 2003.
[32] L. Luo, F. Wu, S. Li, Z. Xiong, and Z. Zhuang, "Advanced motion threading for 3-D wavelet video coding," *Signal Process. Image Commun.*, vol. 19, no. 7, pp. 601–616, Aug. 2004.
[33] G. Feideropoulou and B. Pesquet-Popescu, "Stochastic modelling of the spatio-temporal wavelet coefficients: Application to quality enhancement and error concealment," *EURASIP J. Signal Process. Appl.*, no. 12, pp. 1931–1942, Sep. 2004.
[34] M. Dai, D. Loguinov, and H. Radha, "Rate-distortion modeling of scalable video coders," in *IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2004.
[35] J. Ohm, M. van der Schaar, and J. W. Woods, "Interframe wavelet coding-motion picture representation for universal scalability," *Image Commun. (Special Issue on Digital Cinema)*, 2004.
[36] J. R. Ohm, *Multimedia Communication Technology*. Berlin, Germany: Springer, 2004.

Mingshi Wang received the B.S. degree from the Department of Electrical Engineering and Information Science, University of Science and Technology of China, Hefei, in 2001 and the M.S. degree from the Department of Electrical and Computer Engineering, University of California, Davis, in 2003, where he is currently pursuing the Ph.D. degree.

His research interests include medical imaging, optical imaging system design, and statistical signal processing.

Mihaela van der Schaar (SM'04) received the M.S. and Ph.D. degrees from Eindhoven University of Technology, Eindhoven, The Netherlands, in 1996 and 2001, respectively.

Prior to joining the Electrical Engineering Department, University of California, Los Angeles, in 2005, between 1996 and 2003 she was a Senior Researcher with Philips Research in The Netherlands and the United States, where she led a team of researchers working on multimedia coding, processing, networking, and streaming algorithms and architectures. From January to September 2003, she was also an Adjunct Assistant Professor at Columbia University. From July 2003 until July 2005, she was an Assistant Professor in the Electrical and Computer Engineering Department, University of California, Davis. She has published extensively on multimedia compression, processing, communications, networking, and architectures. She has received 22 U.S. patents with several pending. Since 1999, she has been an active participant in the ISO Motion Picture Expert Group (MPEG) standard to which she made more than 50 contributions and for which she received two ISO recognition awards. She also chaired for three years the ad hoc group on MPEG-21 Scalable Video Coding and cochaired the MPEG ad hoc group on Multimedia Test-bed. She was a Guest Editor of the *EURASIP Journal on Signal Processing Applications* Special Issue on Multimedia over IP and Wireless Networks and General Chair of the Picture Coding Symposium 2004, the oldest conference on image/video coding. She was an Associate Editor of *SPIE Electronic Imaging Journal*.

Prof. van der Schaar has been a Member of the Technical Committee on Multimedia Signal Processing of the IEEE Signal Processing Society. She was an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA. Currently, she is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and an Associate Editor of IEEE SIGNAL PROCESSING LETTERS.