Rate and Distortion Modeling of CGS Coded Scalable Video Content

Hassan Mansour, *Member, IEEE,* Panos Nasiopoulos, *Member, IEEE,*
and Vikram Krishnamurthy, *Fellow, IEEE*

Abstract

In this paper, we derive single layer and scalable video rate and distortion models for video bitstreams encoded using the Coarse Grain Quality Scalability (CGS) feature of the scalable extension of H.264/AVC. In these models, we assume the source is Laplacian distributed and compensate for errors in the distribution assumption by linearly scaling the Laplacian parameter λ. Moreover, we present simplified approximations of the derived models that allow for a run-time calculation of sequence dependent model constants. Our models use the mean absolute difference (MAD) of the prediction residual signal and the encoder quantization parameter (QP) as input parameters. Consequently, we are able to estimate the residual MAD, bitrate, and distortion of a future video frame at any QP value and for both base-layer and CGS layer packets. We also present simulation results that demonstrate the accuracy of the proposed models.

Index Terms

Rate-distortion modeling, scalable video coding, coarse grain scalability.

I. Introduction

Video rate and distortion modeling lies at the core of model-based coder control algorithms, whereby a coded video packet is abstracted in terms of its rate R_s and the distortion D_s of the resulting coded picture [1]. The rate is defined as the size in bytes of the video packet and the distortion is often measured in terms of the mean-square-error (MSE) or peak-signal-to-noise-ratio (PSNR) between the uncoded original and the reconstructed picture.

Rate-distortion (R-D) models are used in real-time coder and transmission control algorithms to predict the rate and distortion of a video packet prior to the completion of the encoding process. However, scalability and encoder complexity add a variety of parameters that render traditional rate and distortion models inaccurate. For instance, the interdependency between rate-distortion optimized (RDO) motion estimation and rate-control in H.264/AVC is described as a "chicken and egg" dilemma [2]. Fig. 1 shows a simplified block diagram of a video encoder with

H. Mansour is with the Departments of Mathematics and Computer Science at the University of British Columbia, 2356 Main Mall, Vancouver, BC Canada V6T 1Z4. Email: hassanm@cs.ubc.ca.

P. Nasiopoulos, and V. Krishnamurthy are with the Department of Electrical and Computer Engineering at the University of British Columbia, 2356 Main Mall, Vancouver, BC Canada V6T 1Z4. Email: {panosn, vikramk}@ece.ubc.ca Phone: 1-604-781-9967. Fax: 1-604-822-9013.

two possible parameter extraction points, the first right after the prediction stage, and the second after transform and quantization.

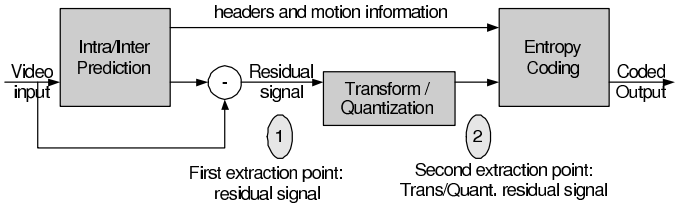


Fig. 1. Simplified encoder block diagram showing the possible extraction points for rate-distortion model parameters.

A more pressing complication arises from the development of scalability features in the state of the art scalable video coding standard. In SVC, quality scalability is achieved using medium grain or coarse grain scalability (MGS/CGS) that deliver quality refinements to a preceding layer representation. A CGS enhancement layer is composed of packets that contain the re-quantized transform coefficients of the residual signal using a smaller quantization step size. When utilizing inter-layer prediction, only the quantization refinements are encoded in the CGS enhancement layer packets [3]. Every CGS layer can be partitioned into several MGS enhancement sub-layers. For example, suppose the CGS layer is split into three MGS layers according to the MGS vector [3 3 10]. Then, the 16 transform coefficients in every 4x4 CGS transform block are split into three MGS groups, such that MGS layer 0 contains the DC coefficient and 2 lowest frequency AC coefficients, MGS layer 1 contains the next 3 low frequency AC coefficients, and MGS layer 2 contains the remaining transform coefficients [3]. These new features in SVC result in a divergence from existing rate-distortion models, calling for the development of improved models that can accurately capture the runtime rate-distortion behavior.

A. Main Results

The main results of this paper can be summarized as follows:

1) We develop analytical rate and distortion models that capture the dynamic behavior of single layer and scalable video coding using the CGS feature in SVC. We assume in these models that the DCT coefficients follow a Laplacian distribution[1].
2) The proposed models can be used to estimate the rate and distortion behavior of future video frames since the model parameters are based on the ℓ_1 norm of the prediction residual. The ℓ_1 norm of the prediction residual is the maximum likelihood estimator of the Laplacian parameter λ and can be updated by linear regression.
3) If the true DCT coefficient distribution diverges from the Laplacian assumption, we propose linear scaling of the Laplacian parameter λ to compensate for the error in the assumption and show that such scaling results in an accurate estimate of the rate and distortion behavior.

[1]While using the Laplacian function to estimate the DCT coefficient distribution is not new, our contribution lies in finding the correct quantization step size and associated Laplacian parameter from the integer transform coefficients used in H.264/AVC. Moreover, we extend this model for the scalable case and derive closed form solutions for the rate and distortion expressions.

4) We show that the analytically derived models can be approximated using simplified empirical models which we have derived in [4].

Remark: While the rate and distortion models derived in this paper correspond to the CGS scalability feature in SVC, it is possible to extend these models for medium grain quality scalability (MGS). The use of MGS coded video bitstreams results in enhancement sub-layers with a linear rate-PSNR relationship within a single CGS layer as illustrated in the example in Fig. 2. This linear relationship was observed for all tested video sequences, however, a detailed analysis is outside the scope of this work.

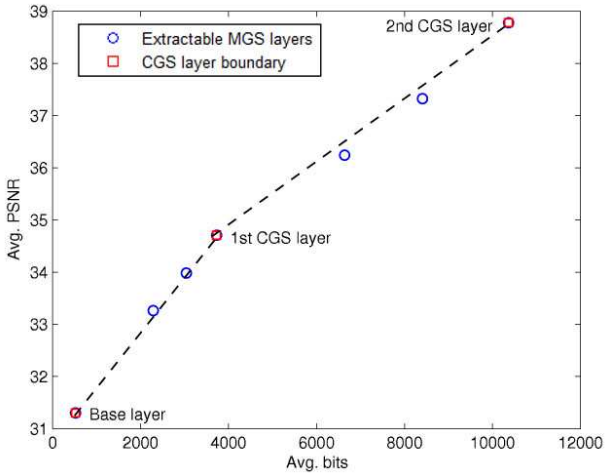


Fig. 2. Example of the operational R-D points achievable with MGS scalability. Foreman sequence is encoded at base-layer QP = 38, and two CGS enhancement layers at QP = 32 and 26. Each CGS layer is divided into three MGS layers. The rate-PSNR behavior is linear within a single CGS layer.

The remainder of this paper is organized as follows. We present in Section II a detailed review of video rate and distortion models. In Section III, we analyze the transform and quantization stages employed in H.264/AVC. Next we develop analytical rate and distortion models in Section IV. These models capture the rate-distortion behavior of CGS coded enhancement layers in scalable video bitstreams. Section V presents simplified empirical models that approximate the performance of the analytical derivations. Performance evaluations are presented in Section VI, and we finally draw our conclusions in Section VII.

II. Literature Review

Many video rate and distortion models have been proposed in the literature. While these models are useful, they do not exploit various aspects of scalable video coding.

A. Generalized Rate-Distortion Models

One of the earliest and more accurate video rate-distortion models [5] devises an empirical formula that relates the video packet bit-rate R_s and its distortion D_s. This relationship is expressed as follows:

$$D_s(R_s) = D_o + \frac{\theta_o}{R_s - R_o} \tag{1}$$

TABLE I
GLOSSARY OF ACRONYMS AND VARIABLES

CGS	Coarse Grain Quality Scalability
DCT	Discrete Cosine Transform
$\mathbf{E}$	4x4 matrix of 1s
MAD	Mean Absolute Difference
MGS	Medium Grain Quality Scalability
ML	Maximum Likelihood
MSE	Mean Squared Error
PSNR	Peak Signal to Noise Ratio
QP	Quantization Parameter
RMSE	Root Mean Squared Error
SATD	Sum of Absolute Transform Differences
SVC	Scalable Video Coding
$f_L(.)$	Laplacian density function
f	Rounding factor that controls the quantization near zero
p	Quantization parameter variable
q	Quantization step size variable
λ	Laplacian distribution parameter

where D_o, θ_o, and R_o are model parameters that depend on the video content and encoder. A major advantage of the model shown in (1) is that it expresses the video distortion as a convex function of the bit-rate.

Another rate distortion model presented in [6] expresses the rate-distortion relationship as an exponential function scaled by the variance of the transform coefficients. This model can be expressed as follows:

$$D_s(R_s) = \sigma^2 e^{-aR_s}, \tag{2}$$

where a is a sequence dependent parameter, and σ^2 is the variance of the transform coefficients.

B. Real-time Rate-Distortion Models

Several models have been proposed to predict the rate and distortion of a video packet prior to the completion of the encoding process In [7], simplified MPEG-2 rate control models are used to estimate the rate and distortion of coded video frames. These models are given by the mean absolute difference (MAD) between every two successive frames ζ_M and the quantization step size Q_s as shown below:

$$D(Q) = aQ_s, \quad R(\zeta_M, Q_s) = b\frac{\zeta_M}{Q_s}, \tag{3}$$

where a and b are model parameters derived empirically for each sequence.

C. Distribution-based Rate-Distortion models

For more accurate representations of the video rate-distortion behavior, several models have been proposed based on the q-domain and the ρ-domain of the residual frame I_r [1], [8]–[10]. The q-domain refers to modeling the

source rate and distortion as a function of the quantization step size and the complexity of the residual signal based on a statistical fit to the distribution of residual samples. In the ρ-domain the rate and distortion are modeled as a function of ρ, where ρ refers to the percentage of zero transform coefficients. The most common distributions used in the above models are the Laplacian and the Cauchy distributions with density functions $f_L(x,\lambda)$ and $f_C(x,\mu)$, respectively, shown below:

$$f_L(x,\lambda) = \tfrac{1}{2\lambda}\exp\{\tfrac{-|x|}{\lambda}\}, \quad f_C(x,\mu) = \tfrac{1}{\pi}\tfrac{\mu}{\mu^2+x^2}, \tag{4}$$

where λ and μ are the respective distribution parameters.

In [11], rate and distortion models are developed based on the ρ-domain information for MPEG-4 and H.263 coded video content at the macroblock level for rate-control applications. The model assumes a Laplacian distribution of DCT coefficients, expresses the rate as a linear function of ρ and uses the exponential relationship between rate and distortion shown in (2). These models are shown below:

$$\begin{aligned} R(\rho) &= \theta(1-\rho), \\ D(\rho) &= \sigma^2 e^{-a(1-\rho)}, \end{aligned} \tag{5}$$

where θ and a are sequence dependent constants, and σ^2 is the distribution variance.

In [1], the rate and distortion models are based on approximations of the ρ-domain behavior and written as the following functions of the quantization step size and sum of absolute transform differences (SATD) of the residual data:

$$\begin{aligned} &R_s(Q_s, \mathrm{SATD}(Q_s)) = \alpha\tfrac{\mathrm{SATD}(Q_s)}{Q_s^{p_1}}, \\ &D_s(Q_s, \mathrm{SATD}(Q_s)) = \beta\mathrm{SATD}(Q_s)Q_s^{p_2} + D_{\mathrm{SKIP}}(Q_s), \end{aligned} \tag{6}$$

where $D_{\mathrm{SKIP}}(Q_s)$ is the distortion of the SKIP mode macroblocks, and α and β are sequence dependent model parameters. The exponentials p_1 and p_2 are equal to 1 for P and B frames, and equal to 0.8 and 1.2, respectively, for I frames. The disadvantage of this model lies in the parameter extraction which is performed late in the encoding process after transform and quantization. Therefore, in order to select an appropriate quantization parameter for the target application, the encoder will have to perform several iterations of prediction, transform and quantization. This process incurs a coding delay which can severely degrade the performance of real-time coder control applications.

In [8], video rate and distortion models are derived for H.264/AVC encoded video streams. The models are derived assuming that the residual source samples follow a Laplacian distribution. The distortion is calculated based on the q-domain analysis which computes the error in quantizing a Laplacian distributed source signal. The bit-rate is modeled as a linear combination of the number of non-zero quantized transform coefficients N_{nz} and the ℓ_1 norm of the quantized transform coefficients E_{QTC}. These two parameters are estimated using the statistical distribution

$f_L(x,\lambda)$ of residual samples. The proposed rate and distortion models are therefore expressed as follows:

$$\begin{aligned} R_s(Q_s,\lambda) &= \alpha N_{\mathrm{nz}} + \beta E_{\mathrm{QTC}}, \\ D_s(Q_s,\lambda) &= 2\int_{0}^{\frac{5}{6}Q_s} x^2 f_L(x,\lambda)dx \\ &\quad +2\sum_{i=1}^{\infty}\int_{(i-\frac{1}{6})Q_s}^{(i+\frac{5}{6})Q_s} (x-iQ_s)^2 f_L(x,\lambda)dx, \end{aligned} \tag{7}$$

where α and β are scaling constants, and λ is the Laplacian distribution parameter that can be calculated from the sample variance: $\sigma_x^2 = 2\lambda^2$. Moreover, the authors derive simplified transform domain equations by relating the distribution of the spatial domain residual samples with their transformed coefficients. This is performed by assuming that the source samples in a 4x4 block are Markovian, and using the correlation coefficient of the spatial domain samples to calculate the transform domain sample variance. However, the work makes several simplifying assumptions especially in the rounding error made during quantization and they do not address SKIP mode[2] macroblocks which occur very frequently in H.264/AVC encoding.

The work in [9] is similar to the above mentioned work with two main distinctions: the transform coefficients of the residual source samples are assumed to follow a Cauchy distribution $f_C(x,\mu)$, and the bit-rate is estimated using the statistical entropy $H(Q_s)$ of the quantized coefficients. The derived rate and distortion models and their simplified versions proposed in this work are shown below:

$$\begin{aligned} &R_s(Q_s,\mu) = H(Q_s) = -\sum_{i=-\infty}^{\infty} P(iQ_s)\log_2\left(P(iQ_s)\right), \\ &\text{where } P(iQ_s) = \int_{(i-\frac{1}{2})Q_s}^{(i+\frac{1}{2})Q_s} f_C(x,\mu)dx, \\ &D_s(Q_s,\mu) = \sum_{i=-\infty}^{\infty}\int_{(i-\frac{1}{2})Q_s}^{(i+\frac{1}{2})Q_s} \left(x-iQ_s\right)^2 f_C(x,\mu)dx. \\ &\text{Simplified models: } R_s(Q_s) \approx aQ_s^{-\alpha}, \quad D_s(Q_s) \approx bQ_s^{\beta}, \end{aligned} \tag{8}$$

where a, b, α, and β are sequence dependent parameters. However, several simplifying assumptions are made in terms of the quantization rounding error, the models are based strictly on the distribution of transform coefficients, and the SKIP mode is not addressed in the derivation.

Finally, a scalable video rate-distortion model is derived [10] for progressive fine granular scalable (PFGS) video coders. The model assumes that the residual transform coefficients follow a sum of Laplacian distributions and computes the distortion as the error in quantizing such sources. The rate is derived following the ρ-domain analysis in which it can be written as a linear function of the percentage of non-zero transform coefficients. This results in the following rate-distortion behavior:

$$D_s(R_s) = \sigma_x^2 - (a\log^2 R_s + b\log R_s + c)R_s, \tag{9}$$

[2]The SKIP mode is an Inter-prediction mode for P-frame macroblocks in which no texture information is encoded. Therefore, the resulting SKIP mode distortion is not that of a quantization error but it is equal to the energy of the residual signal.

where a, b, and c are sequence dependent model parameters, and σ_{x^2} is the variance of the transform coefficients. This scalability model is similar to the type supported by MPEG-4 [12] and the early version of quality scalability that was supported in SVC prior to its exclusion from the standard due to the additional complexity imposed on the decoder.

While these models are faithful to the assumptions taken, several new features in SVC result in a divergence from existing rate-distortion models, calling for the development of improved models that can accurately capture the runtime rate-distortion behavior. These rate and distortion models can be used in conjunction with rate control and stochastic system models to derive transmission control policies that address the requirements and constraints of wireless video transmission applications.

III. Transform and Quantization in H.264/AVC

In this section, we derive expressions that relate the integer transform coefficients and associated quantization step sizes to their corresponding DCT coefficients and a uniform quantization step size.

A. Transform

The purpose of introducing a transform is to reduce the spatial correlation to improve compression. Since the smallest prediction block employed in H.264/AVC is of size 4x4 pixels, the standard uses a 4x4 transform block size which also leads to a significant reduction in ringing artifacts compared to the 8x8 transform size used in previous standards.

The discrete cosine transform has been the transform of choice in previous video coding standards. This is due to its close approximation of the statistically optimal Karhunen-Loève transform for image signals [13]. Moreover, given an $N\mathrm{x}N$ signal matrix $\mathbf{x}$, the DCT coefficients $\hat{\mathbf{x}}$ can be computed as the linear transformation $\hat{\mathbf{x}} = \mathbf{F}\mathbf{x}\mathbf{F}^T$, where $\mathbf{F}$ is the forward transform matrix with the kth row, nth column element F_{kn} given by:

$$F_{kn} = c_k \sqrt{\frac{2}{N}} \cos\left[\left(n + \frac{1}{2}\right)\frac{k\pi}{N}\right], \tag{10}$$

where $c_0 = \sqrt{2}$, and $c_k = 1$ for $k > 0$. However, the DCT entries are irrational numbers that require a floating point representation which is platform dependent. Therefore, the transform used in H.264/AVC is an integer approximation of the 4x4 DCT achieved by rounding scaled entries of the DCT matrix to the nearest integer as shown below:

$$\mathbf{H} = \mathrm{round}(2.5 \times \mathbf{F}). \tag{11}$$

Therefore, the forward integer transform matrix $\mathbf{H}$ and scaled inverse transform matrix $\mathbf{H_i}$ are given as follows

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix}, \quad \mathbf{H_i} = \begin{bmatrix} 1 & 1 & 1 & 1/2 \\ 1 & 1/2 & -1 & -1 \\ 1 & -1/2 & -1 & 1 \\ 1 & -1 & 1 & -1/2 \end{bmatrix}, \tag{12}$$

which results in the relationship

$$\mathbf{H_i D^{-1} H} = \mathbf{I}, \tag{13}$$

where $\mathbf{D}$ is a diagonal matrix given by $\mathbf{D} = \text{diag}\{4, 5, 4, 5\}$.

Since the integer transform and its scaled inverse are non-orthonormal matrices, a 4x4 block $\mathbf{x}$ can only be reconstructed by first scaling the transformed coefficients $\hat{\mathbf{x}}$ with a matrix $\mathbf{V}$ before applying the inverse transform. This can be illustrated as follows:

$$\begin{aligned} \hat{\mathbf{x}} &= \mathbf{H x H^T}, \\ \mathbf{x} &= \mathbf{H_i} \left(\hat{\mathbf{x}} \oslash \mathbf{V} \right) \mathbf{H_i^T}, \end{aligned} \tag{14}$$

where $\oslash$ indicates element by element division, and $\mathbf{V}$ is given in (19).

B. Quantization

Let $\mathbf{X}$ be a 4x4 block from the residual frame I_r, and let $\hat{\mathbf{X}}$ be the corresponding transformed block. The H.264/AVC encoder performs a scalar quantization of the coefficients $\hat{\mathbf{X}}$ using only integer multiplication and right bit-shifts to avoid any division [13].

Let $\hat{\mathbf{X}}_\mathbf{q}$ and $\hat{\mathbf{X}}_\mathbf{r}$ be the quantized and reconstructed blocks corresponding to $\mathbf{X}$, respectively. A typical scalar quantization and reconstruction process can be performed by

$$\hat{\mathbf{X}}_\mathbf{q} = \text{sign}(\hat{\mathbf{X}}) \left\lfloor \frac{|\hat{\mathbf{X}}|}{Q_s} + f \right\rfloor, \quad \text{and} \quad \hat{\mathbf{X}}_\mathbf{r} = \hat{\mathbf{X}}_\mathbf{q} Q_s, \tag{15}$$

where Q_s is the quantization step, and $f \in [0, 0.5]$ is a rounding factor that controls the quantization near zero. *Note that in the above equation and in the remainder of this paper, the addition of a scalar to a matrix denotes adding the scalar to each element of the matrix.*

In order to reduce complexity and avoid division operations, the H.264/AVC encoder implements the quantization process by multiplying the transformed signal by a scalar multiplier extracted from periodic quantization followed by right bit-shifts to reduce the dynamic range of the quantized coefficients. The standard defines a quantization parameter (QP) which is used to select the multiplier from the quantization tables. Let $Q \in \{0, 1, \cdots 51\}$ be the encoding QP, the quantization and reconstruction operations can be summarized as follows

$$\begin{aligned} \hat{\mathbf{X}}_\mathbf{q} &= \text{sign}(\hat{\mathbf{X}}) \left\lfloor \left(|\hat{\mathbf{X}}| \,.\, M(Q_m) + f 2^{15+Q_e} \right) 2^{-(15+Q_e)} \right\rfloor, \\ \hat{\mathbf{X}}_\mathbf{r} &= \hat{\mathbf{X}}_\mathbf{q} \,.\, S(Q_m) 2^{Q_e}, \end{aligned} \tag{16}$$

where $.$ is the Hadamard product which denotes elementwise multiplication, $Q_m = Q \bmod 6$, $Q_e = \lfloor Q/6 \rfloor$, $M(Q_m)$ is a 4x4 forward scaling matrix, and $S(Q_m)$ is the reconstruction scaling matrix. The scaling matrices $M(Q_m)$ and $S(Q_m)$ are derived from standard defined quantization tables [14], and have the following relationship

$$M(Q_m) \,.\, S(Q_m) \,.\, \mathbf{V} \cong 2^{21}, \tag{17}$$

where $\mathbf{V}$ is the matrix defined in (19).

Finally, the reconstructed residual signal $\mathbf{X}_r$ is calculated by performing the inverse transform and rounding operations shown below

$$\mathbf{X}_r = \left\lfloor \left(\mathbf{H_i} \hat{\mathbf{X}}_\mathbf{r} \mathbf{H_i^T} + 2^5 \right) 2^{-6} \right\rfloor . \tag{18}$$

C. Analysis of the Transform and Quantization Stages

Given the scaling matrices $M(Q_m)$ and $S(Q_m)$ which are derived from standard defined quantization tables [14], the following relation holds:

$$M(Q_m) \,.\, S(Q_m) \,.\, \mathbf{V} \cong 2^{21},$$

where $\mathbf{V}$ is the matrix given by

$$\mathbf{V} = \mathbf{D}\mathbf{E}\mathbf{D}^\mathbf{T} = \begin{bmatrix} 16 & 20 & 16 & 20 \\ 20 & 25 & 20 & 25 \\ 16 & 20 & 16 & 20 \\ 20 & 25 & 20 & 25 \end{bmatrix}, \tag{19}$$

where $\mathbf{E}$ is a 4x4 matrix of 1s.

Consequently, we get the following quantization and true reconstruction equations

$$\begin{aligned} \hat{\mathbf{X}}_\mathbf{q} &= \mathrm{sign}(\hat{\mathbf{X}}) \left\lfloor \frac{|\hat{\mathbf{X}}|}{\mathbf{V}.S(Q_m)2^{(Q_e-6)}} + f \right\rfloor , \\ \tilde{\hat{\mathbf{X}}} &= \hat{\mathbf{X}}_\mathbf{q} \,.\, \mathbf{V} S(Q_m) 2^{Q_e-6}, \end{aligned} \tag{20}$$

where $\tilde{\hat{\mathbf{X}}}$ is the true reconstruction of the quantized coefficient block $\hat{\mathbf{X}}_\mathbf{q}$. It can be seen from (16), (20), and (18) that

$$\tilde{\hat{\mathbf{X}}} = \hat{\mathbf{X}}_\mathbf{r} \,.\, \mathbf{V} 2^{-6}, \tag{21}$$

which is a natural relationship since $\mathbf{H_i}$ is not the true inverse but a scaled inverse of the forward transform $\mathbf{H}$.

Consequently, (20) shows that the quantization step size Q_s of Eq. (15) employed in the H.264/AVC quantization process is a 4x4 matrix given by

$$Q_s = \mathbf{V} \,.\, S(Q_m) 2^{Q_e-6}, \tag{22}$$

To better illustrate the quantization operation, we show the quantization step matrix for a quantization parameter value $Q = 26$:

$$Q_s = \begin{bmatrix} 52 & 80 & 52 & 80 \\ 80 & 125 & 80 & 125 \\ 52 & 80 & 52 & 80 \\ 80 & 125 & 80 & 125 \end{bmatrix}. \tag{23}$$

It can be seen from above that different transform subbands are subjected to different quantization step sizes. This is due to the non-uniform scaling that is required to enable the integer transform used in H.264/AVC. Moreover, the integer transform coefficient distribution varies with the location of the subband in the 4x4 transform block. As

a result, any rate/distortion model has to treat each subband separately, which is a cumbersome approach employed by [8]. Alternatively, we can convert the integer transform coefficients to DCT coefficients using the block scaling process described below and find a single quantization step size that is applied to all the DCT subbands. We will show in subsequent sections that this allows us to model the rate and distortion behaviour using a linearly scaled version of the pixel-domain residual distribution.

In [15], the authors find a scaling matrix $\mathbf{W}$ which compensates for the different norms of the forward and scaled-inverse integer transforms. To derive the matrix $\mathbf{W}$, let $\mathbf{G} = \mathbf{H^{-1}}^{\mathbf{T}}\mathbf{H^{-1}}$, then $\mathbf{W}$ is given by:

$$\mathbf{W} = \sqrt{\mathbf{GEG^T}} = \begin{bmatrix} 1/4 & 1/\sqrt{40} & 1/4 & 1/\sqrt{40} \\ 1/\sqrt{40} & 1/10 & 1/\sqrt{40} & 1/10 \\ 1/4 & 1/\sqrt{40} & 1/4 & 1/\sqrt{40} \\ 1/\sqrt{40} & 1/10 & 1/\sqrt{40} & 1/10 \end{bmatrix}, \tag{24}$$

where $\mathbf{E}$ is a 4x4 matrix of 1s.

Consequently, we find the corresponding DCT coefficients X_w and associated uniform quantization step size Q_w by multiplying the integer transform and quantization matrices with $\mathbf{W}$ as shown below:

$$\begin{aligned} X_w &= X \,.\, \mathbf{W}, \\ Q_w &= Q_s \,.\, \mathbf{W} = q_w \mathbf{E}, \end{aligned} \tag{25}$$

where q_w is a scalar and $\mathbf{E}$ is the 4x4 matrix of 1s.

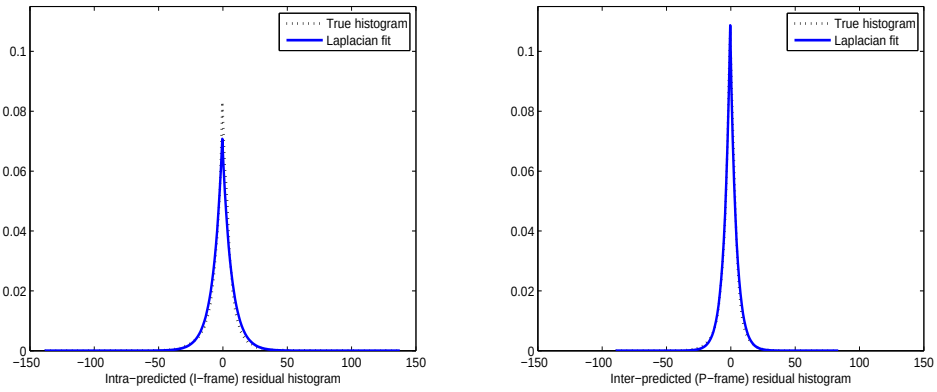


Fig. 3. Illustration of the pixel domain residual histogram and the Laplacian distribution fit for selected Intra- and Inter- coded frames from the Foreman sequence with CIF resolution. The plots are accumulated over 7 I-frames and 192 P-frames at QP = 38.

IV. Analytical Derivation of the Rate-Distortion Behavior

In this section, we present an approach to rate-distortion modeling based on the statistical distributions of the residual signal and its transform coefficients. We derive the models based on the analysis of the transform and quantization stages for H.264/AVC presented in Section III.

The objective of a video encoder is to represent a sequence of images in as few bits as possible while maintaining a desired reconstruction quality. Modern video encoders achieve this task through four stages depicted in Fig 1. The first stage is Intra/Inter prediction which captures the spatial and temporal correlations resulting in a pixel domain residual signal. The second stage transforms the residual signal in order to collect the energy of the signal in fewer samples than the pixel domain residual signal, resulting in a sparser representation of the residual. The third stage quantizes the transform coefficients to reduce their dynamic range. Finally, a fourth stage translates the quantized coefficients into a binary bitstream through entropy coding.

A. Distribution of Residual Samples

As described in Sec.II, several models have been presented in the literature that attempt to model the residual transform coefficients [1], [6], [8], [9]. These include the generalized Gaussian, Laplacian, and Cauchy distributions. The most widely used distribution is the Laplacian which we will adopt in this work for its accuracy, simplicity, and analytical tractability. Moreover, our experiments have shown that the distribution of pixel domain residual samples can also be approximated by the Laplacian distribution, an observation supported by [8].

Let $f_X(x, \lambda)$ be the zero-mean Laplacian density function given by

$$f_X(x, \lambda) = \frac{1}{2\lambda} \exp\left(\frac{-|x|}{\lambda}\right), \tag{26}$$

where λ parameterizes the Laplacian distribution. Given a sample data set $\mathcal{X}$ of size N, where $x_i \in \mathcal{X}, i = 1, 2, \cdots N$, the maximum-likelihood (ML) estimator of λ is given by

$$\lambda = \frac{1}{N} \sum_{i=1}^{N} |x_i|. \tag{27}$$

It is easy to see from (27) that the ML estimator of λ is equal to the arithmetic mean of the residual samples. This measure also corresponds to the mean absolute difference (MAD) between the image pixels and the motion compensated prediction in the case of the pixel domain samples and the sum of absolute transform difference (SATD) in the case of the transform domain samples.

Fig. 3 demonstrates the accuracy of the ML Laplacian fit to the histogram of selected Intra- and Inter- coded frames from the Foreman sequence. Moreover, using the MATLAB implementation of the two sample Kolmogorov-Smirnov test (ks-test) [16], we compare the similarities of the residual frame distributions with the Laplacian distribution. Out of 2000 coded video frames, 79.64% of the time the Kolmogorov-Smirnov test requirements were satisfied with an average P-value of 0.2932. The P-value compares the probability that a sample that is drawn from the Laplacian population deviates from the Laplacian distribution as much as the samples in the residual frame. Therefore, a larger P-value is better. The null hypothesis in our case is that the residual samples are drawn from a Laplacian distrubtion. The ks-test was successful almost 80% of the time at the 5% significance level and the P-value is not small enough to reject this claim. Thus the hypothesis that the distribution of residual samples follows a Laplacian distribution cannot be rejected.

B. Base-layer Models

We start by deriving the rate and distortion models for base-layer frames encoded using an H.264/AVC compliant encoder.

1) Distortion Model: Let X be the pixel domain video sample vector. The video distortion D_s of a video frame is measured in terms of the MSE between the original picture and its reconstruction $\tilde{X}$ in the pixel domain. This can be expressed as the expected value of the ℓ_2 norm of the encoding noise as shown below:

$$D_s = \mathbb{E}\left[\|X - \tilde{X}\|_2^2\right]. \tag{28}$$

The samples in a coded video frame are divided into macroblocks compressed using transform coding, and SKIP mode macroblocks in which the prediction residual is simply set to zero. Therefore, the distortion term above is separable into the two expectation terms shown below:

$$\begin{aligned} D_s &= \mathbb{E}\left[\|X_c - \tilde{X}_c\|_2^2\right] + \mathbb{E}\left[\|X_{\text{SKIP}} - \tilde{X}_{\text{SKIP}}\|_2^2\right] \\ &= \mathbb{E}\left[\|X_c - \mathcal{Q}(X_c - \mathcal{P}(X_c)) - \mathcal{P}(X_c)\|_2^2\right] \\ &\quad + \mathbb{E}\left[\|X_{\text{SKIP}} - \mathcal{P}(X_{\text{SKIP}})\|_2^2\right] \\ &= D_c + D_{\text{SKIP}}, \end{aligned} \tag{29}$$

where $\mathcal{Q}(.)$ combines the processes of transform-quantization-inverse_transform-inverse_quantization, $\mathcal{P}(.)$ is the prediction process, and the subscripts c and SKIP correspond to the transform domain samples and SKIP macroblock samples, respectively. We use the terms D_c and D_{SKIP} to label the distortion due to quantizing the transformed residual samples, and the distortion due to SKIP macroblocks, respectively.

In what follows we estimate the distortion term D_c based on the Laplacian assumption for the residual sample distribution. The term D_{SKIP} can also be found statistically, however, since we base our analysis on the availability of the residual samples, it is more reasonable and accurate to calculate D_{SKIP} directly.

The distortion D_c results from the quantization error and associated integer rounding operations performed in the transform domain. To estimate the transform domain distortion $\hat{D}_c$, we compute the q-domain error of quantizing a Laplacian distributed source signal with a quantization step matrix Q_s and a dead-zone parameter f. This error is estimated by the expected value of the ℓ_2 norm of the quantization noise $(\hat{X} - \tilde{\hat{X}})$.

Assuming that every corresponding DCT coefficient $\hat{\mathbf{Y}}(u, v)$ is sampled from a Laplacian distribution with Laplacian parameter $\hat{\lambda}$, the quantization error $D_c(Q_w, \hat{\lambda})$ is given by the following expression:

$$\begin{aligned} \hat{D}_c(Q_w, \hat{\lambda}) &= \mathbb{E}\left[\| \left(\hat{X} - \tilde{\hat{X}}\right) \cdot W\|_2^2\right] \\ &= \sum_{i=-\infty}^{\infty} \int_{(i-f)Q_w}^{(i+1-f)Q_w} (\hat{Y} - iQ_w)^2 f_{\hat{Y}}(y, \hat{\lambda}) dy, \end{aligned} \tag{30}$$

which can be written as a function of the scalar q_w:

$$\begin{aligned}\hat{D}_c(q_w,\hat{\lambda}) &= 2\int_0^{(1-f)q_w} y^2 f_{\hat{Y}}(y,\hat{\lambda})dy \\ &+ 2\sum_{i=1}^{\infty}\int_{(i-f)q_w}^{(i+1-f)q_w}(y-iq_w)^2 f_{\hat{Y}}(y,\hat{\lambda})dy, \\ &= 2\hat{\lambda}^2 + \left[(2f-1)q_w^2 - 2\hat{\lambda}q_w\right]\frac{e^{-(1-f)q_w/\hat{\lambda}}}{1-e^{-q_w/\hat{\lambda}}},\end{aligned} \tag{31}$$

where $i = \left\lfloor \frac{\hat{X}}{Q_s} + f \right\rfloor = \left\lfloor \frac{\hat{X}.W}{Q_s.W} + f \right\rfloor$, $\hat{Y} = \hat{X}\,.\,W$ are the newly scaled transform coefficients, $Q_w = Q_s\,.\,W$ is a uniform quantization matrix, and $f_{\hat{Y}}(y,\hat{\lambda})$ is the density function of the scaled transform coefficients $\hat{Y}$.

Model parameter estimation using pixel domain samples:

The distortion model derived above assumes that the DCT transform coefficients are Laplacian distributed. However, the true distribution of coefficients is closer to the generalized Gaussian distribution with a shape parameter < 1. Therefore, to compensate the error in the Laplacian assumption and estimate the model parameter using pixel domain residual samples, we estimate the Laplacian parameter $\hat{\lambda}$ as a linear (or more precisely affine) function of the pixel domain parameter as follows:

$$\hat{\lambda} = a\lambda_X + b, \tag{32}$$

where λ_X is the pixel-domain Laplacian distribution parameter, and a and b are sequence dependent constants that can be calculated from a limited number of frames and updated using linear regression during the encoding process.

2) Bit-Rate Model: The bit-rate of a coded video frame is equal to the number of bits required to code the header and control information plus the number of bits required to encode the quantized levels of the transform coefficients. Let R_s be the size in bits of a coded video frame. Then for a quantization step size q_w, R_s is expressed as:

$$R_s(q_w) = R_c(q_w) + R_h, \tag{33}$$

where R_c and R_h are the numbers bits required to code the texture and header information, respectively.

In [1], a model is proposed to estimate the video frame header size as a linear function of the number of zero and non-zero motion vectors. Other models have used the CAVLC coding assumptions to make the same estimation. We will not address the header bit-rate estimation problem since existing models are accurate enough. Our focus will be on estimation of the bit-rate of quantized transform coefficients of the residual signal.

Previous works that estimate the rate of encoding the quantized transform coefficients $R_c(q_w)$ have modeled the bit-rate as a linear function of the ℓ_0 and ℓ_1 norms [8] [1]. These norms are calculated using the distribution of integer transform coefficients. In [9], the authors calculate the entropy of the quantized coefficients assuming a Cauchy distributed source. Since H.264/AVC uses a highly sophisticated entropy coder such as CABAC, we will also use the entropy of the quantized transform coefficients to model the residual bit-rate. The distinction in our work lies in our choice of a Laplacian distribution to model the scaled transform coefficients. Moreover, the use of entropy to model the residual bit-rate reduces the number of model parameters to those used to estimate the scaled

transform coefficient, MAD $\hat{\lambda}$.

Let iq_w be the quantized coefficient value with probability $\mathbb{P}(iq_w)$ which is calculated as follows:

$$\begin{aligned}\mathbb{P}(iq_w) &= \int_{(i-f)q_w}^{(i+1-f)q_w} f_{\hat{Y}}(y,\hat{\lambda})dy \\ &= \begin{cases} 1-e^{-(1-f)q_w/\hat{\lambda}}, & i=0 \\ \frac{1}{2}e^{-(|i|-f)q_w/\hat{\lambda}}\left(1-e^{-q_w/\hat{\lambda}}\right), & i=\pm1,\pm2,\cdots \end{cases}\end{aligned} \tag{34}$$

Then the entropy of the quantized coefficients $H(q_w)$ for a quantization step q_w is given by

$$\begin{aligned}H(q_w) &= -\sum_{i=-\infty}^{\infty}\mathbb{P}(iq_w)\log_2\left(\mathbb{P}(iq_w)\right) \\ &= -(1-e^{-(1-f)q_w/\hat{\lambda}})\log_2(1-e^{-(1-f)q_w/\hat{\lambda}}) \\ &\quad -e^{-(1-f)q_w/\hat{\lambda}}\left[\log_2(1-e^{-q_w/\hat{\lambda}})-1\right. \\ &\quad \left.+\frac{q_w}{\hat{\lambda}\ln 2}\left(f-\frac{1}{1-e^{-q_w/\hat{\lambda}}}\right)\right]\end{aligned} \tag{35}$$

The quantized coefficient entropy $H(q_w)$ determines the number of bits/pixel required to encode a source signal with a distribution $f_{\hat{Y}}(y,\hat{\lambda})$. Therefore, the coded frame size is given by the sum of the coded residual frame bits $R_c q_w$ and the header bits R_h

$$R_s(q_w) = NH(q_w) + R_h, \tag{36}$$

where N is the number of pixels per video frame. Note that the rate estimation using the quantized signal entropy assumes that the entropy encoder employed by the video encoder can meet these requirements. Fortunately, the CABAC entropy coder supported by H.264/AVC meets these requirements. Moreover, the CABAC coder takes advantage of the dependency between adjacent samples, a feature which is not supported in our entropy calculation. The entropy function derived above assumes that the quantized coefficients are coded independently which results in an upper bound on the real frame bit-rate. Therefore, to compensate for the error in this estimation, we further subject the entropy calculation to linear scaling to match the actual coded video frame rate. A simple least squares approach can be used over a small number of video frames, 5 to 10 frames in general. This linear scaling is sequence dependent but remains constant over the duration of each video sequence until a scene change occurs.

C. Enhancement-layer Models

In CGS scalable coding mode, a base layer representation of the video sequence is encoded at a QP value q_1 and an enhancement layer representation is coded at a QP value q_2, such that $q_2 < q_1$. When no refinement motion vectors are encoded in the enhancement layer, then the enhancement layer will only encode a requantization of the same residual signal that was encoded in the base layer. This residual signal is characterized by the un-coded pixel-domain residual samples X and their corresponding transform coefficients $\hat{Y}$.

Let $\tilde{\hat{Y}}_1$ be the reconstructed transform coefficients of the base-layer representation after quantization with a step size $q_{w,1}$. The enhancement layer representation of the same frame contains the quantization of the base layer quantization-noise signal $(\hat{Y} - \tilde{\hat{Y}}_1)$ using the step size $q_{w,2}$, where $q_{w,2} < q_{w,1}$.

1) Distortion Model: Let $\tilde{\hat{Y}}_2$ be the reconstructed transform coefficient value after decoding the base and enhancement layer representations of a video frame. Then the enhancement layer quantization noise $\hat{D}_{c,2}$ is given by the following expression:

$$\begin{aligned}\hat{D}_{c,2}(q_{w,2}, q_{w,1}, \hat{\lambda}) &= \|\hat{Y} - \tilde{\hat{Y}}_2\|_2^2 \\ &= \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \|\hat{Y} - (iq_{w,1} + jq_{w,2})\|_2^2,\end{aligned} \tag{37}$$

where i and j are the quantization levels of the base and enhancement layer representations, respectively.

Assume that the scaled transform coefficients follow the Laplacian distribution $f_{\hat{Y}}(y, \hat{\lambda})$. Then the enhancement layer distortion can be estimated as the expected value of the ℓ_2 norm of the refinement quantization noise shown in (37). This can be calculated as follows:

$$\begin{aligned}\hat{D}_{c,2}(q_{w,2}, q_{w,1}, \hat{\lambda}) &= \sum_{i=-\infty}^{\infty} \mathbb{P}(iq_{w,1}) \\ &\quad \times \sum_{j=-\infty}^{\infty} \mathbb{P}(jq_{w,2}|iq_{w,1})(\hat{Y} - jq_{w,2} - iq_{w,1})^2,\end{aligned} \tag{38}$$

where $\mathbb{P}(iq_{w,1})$ is the base layer quantized coefficient probability given in (34), and $\mathbb{P}(jq_{w,2}|iq_{w,1})$ is the enhancement layer quantized refinement probability conditioned on the base layer quantization and expressed as follows:

$$\mathbb{P}(jq_{w,2}|iq_{w,1}) = \int_{(j-f)q_{w,2}+iq_{w,1}}^{(j+1-f)q_{w,2}+iq_{w,1}} f_{\hat{Y}}(y, \hat{\lambda}) dy. \tag{39}$$

Consequently, the closed form of the enhancement layer distortion is expressed as follows:

$$\hat{D}_{c,2}(q_{w,2}, q_{w,1}, \hat{\lambda}) = \hat{D}_c(q_{w,2}, \hat{\lambda}) \left[1 - e^{-(1-f)q_{w,1}/\hat{\lambda}} \frac{1 - e^{-q_{w,1}/\hat{\lambda}}}{1 - e^{-2q_{w,1}/\hat{\lambda}}}\right], \tag{40}$$

where $\hat{D}_c(q_{w,2}, \hat{\lambda})$ is the single layer distortion derived in (31) for a quantization step size $q_{w,2}$.

2) Bit-Rate Model: Following the analysis for the base-layer bit-rate, the enhancement layer bit-rate is the sum of the enhancement layer header bits plus the number of bits required to encode the residual signal quantization refinement. We estimate the enhancement layer quantization refinement using the conditional entropy $H(q_{w,2}|q_{w,1})$ of the quantized base-layer quantization noise $(\hat{Y} - \tilde{\hat{Y}}_1)$. The combined stream (base layer + CGS) bit-rate can then be written as a function of the joint entropy $H(q_{w,2}, q_{w,1})$ shown below:

$$H(q_{w,2}, q_{w,1}) = H(q_{w,1}) + H(q_{w,2}|q_{w,1}), \tag{41}$$

where $H(q_{w,1})$ is the base layer quantized coefficient entropy derived in (35). In what follows, we derive the conditional entropy term.

$$\begin{aligned} H(q_{w,2}|q_{w,1}) &= -\sum_{i=-\infty}^{\infty} \mathbb{P}(iq_{w,1}) \\ &\quad \times \sum_{j=-\infty}^{\infty} \mathbb{P}(jq_{w,2}|iq_{w,1}) \log_2(\mathbb{P}(jq_{w,2}|iq_{w,1})) \\ &= \frac{1-e^{-(1-f)q_{w,1}/\hat{\lambda}}}{1-e^{-2q_{w,1}/\hat{\lambda}}} \\ &\quad \times \left[H(q_{w,2}) + \frac{q_{w,1}}{\hat{\lambda}\ln 2} \frac{1-e^{-(2-f)q_{w,1}/\hat{\lambda}}}{1-e^{-2q_{w,1}/\hat{\lambda}}}\right], \end{aligned} \tag{42}$$

where $H(q_{w,2})$ is the single layer quantized coefficient entropy in (35) evaluated at the quantization step $q_{w,2}$.

Therefore, the total scalable frame bit-rate can now be written as follows:

$$R_s(q_{w,1}, q_{w,2}) = N\left(H(q_{w,1}) + H(q_{w,2}|q_{w,1})\right) + R_{h,1} + R_{h,2}, \tag{43}$$

where $R_{h,1}$ and $R_{h,2}$ are the base layer and enhancement layer header bits.

V. Empirical Modeling of the Rate-Distortion Behavior

In this section, we propose simplified rate and distortion models that are derived from empirical data estimates. In these models, we use the quantization parameter (QP) and the mean absolute difference (MAD) of the residual signal $\lambda(p_0)$, obtained at extraction point (1) of Fig. 1, given an initial quantization parameter value p_0. These empirical models can accurately estimate the following:

1) The MAD of the residual signal $\lambda(p_t)$ at any target p_t different from the initial parameter p_0. The prediction error $\lambda(p_t)$ is used to estimate the rate and distortion of the coded packet.
2) The decoded picture distortion measured in terms of the luminance PSNR (Y-PSNR) at any target p_t. The distortion D is expressed as a function of the estimated prediction error $\lambda(p_t)$ and the quantization parameter p_t.
3) The coded packet rate measured in terms of the size in bytes of the CGS enhancement layer packets at any target p_t. For base layer packets, the traditional quadratic bit-rate model and its linear approximation used for rate control in [17] remains valid. For CGS enhancement packets, we have found that the rate R is a linear function of the enhancement layer distortion D.

A. Prediction Error Estimation Model

In H.264/AVC and base-layer SVC, the macroblock (MB) prediction mode is chosen after a rate-distortion optimization (RDO) process, such that the chosen mode minimizes the following Lagrangian cost function

$$J(m, q_s) = D(m, q_s) + \kappa(q_s)R(m, q_s), \tag{44}$$

where m is the macroblock coding mode, q_s is the quantization step size, D is the prediction distortion measured in terms of the sum of absolute differences (SAD) or sum of squared differences (SSD) between the original and

prediction signals, R is the number of bits required to encode the MB using the specific mode, and κ is the Lagrange multiplier given as a function of QP [18].

The rate control algorithm used in H.264/AVC allows for the estimation of the prediction, MAD $\tilde{\lambda}$ of a frame using the following first-order auto-regressive model:

$$\tilde{\lambda}_t = \alpha\tilde{\lambda}_{t-1} + \beta, \quad \alpha \leq 1, \tag{45}$$

where $\tilde{\lambda}_{t-1}$ is the actual MAD of the previous frame, and α and β are model parameters that are initialized to 1 and 0, respectively. $\tilde{\lambda}_0$ is initialized to the actual MAD of the first coded video frame. These parameters are updated by linear regression [17].

The model described in (45) is sufficient when both the current and previous frames have the same QP value. However, the QP values of real-time encoded CGS video frames change in both the base and enhancement layers. Let $p_t, p_{t-1} \in \{0, 1, ...51\}$ be the current and previous QP values, respectively. It is evident from (44) that the resulting residual signal, and consequently the residual MAD, cannot be calculated before the completion of the RDO process which is a function of QP. Consequently, encoders would have to perform a new RDO process to update $\lambda(p_{t-1})$ for different values of p_t, which is very costly in terms of computational resources and encoding delay. The relationship that estimates $\tilde{\lambda}(p_t)$ given the residual MAD at an initial p_{t-1} is shown below:

$$\tilde{\lambda}(p_t) = \tilde{\lambda}(p_{t-1})2^{a(p_t - p_{t-1})}, \tag{46}$$

where a is a model parameter typically valued around 0.07 for most sequences. The model shown in (46) allows an encoder to accurately estimate the prediction error at different QP values using only a single prediction run at an initial p_{t-1}. Fig. 4 illustrates the accuracy of this prediction. The estimated $\lambda(p_t)$ is then used in the rate and distortion estimation models. As a result, optimal encoding can be performed using a total of only two prediction runs, the first run at the initial p_{t-1} and the second run at the optimal target p_t.

B. Distortion (PSNR) Model

The Laplacian-based distortion model derived in (31) can be simplified by observing the variation of D_s as a function of λ and QP independently. Our observations have shown that we can estimate the MSE as a linear function of λ and an exponential function of QP as shown in Fig. 5.

As a result, we approximate the distortion measure using the following expression:

$$D_s(\lambda, p) = a\lambda e^{bp+c}, \tag{47}$$

where a, b, and c are sequence dependent constants.

Let p_b and p_e be the base and enhancement layer QP values, respectively, and let $\tilde{\lambda}_b = \tilde{\lambda}(p_b)$ and $\tilde{\lambda}_e = \tilde{\lambda}(p_e)$ be the respective prediction MAD estimates. For an H.264/AVC coded video stream and H.264/AVC compliant

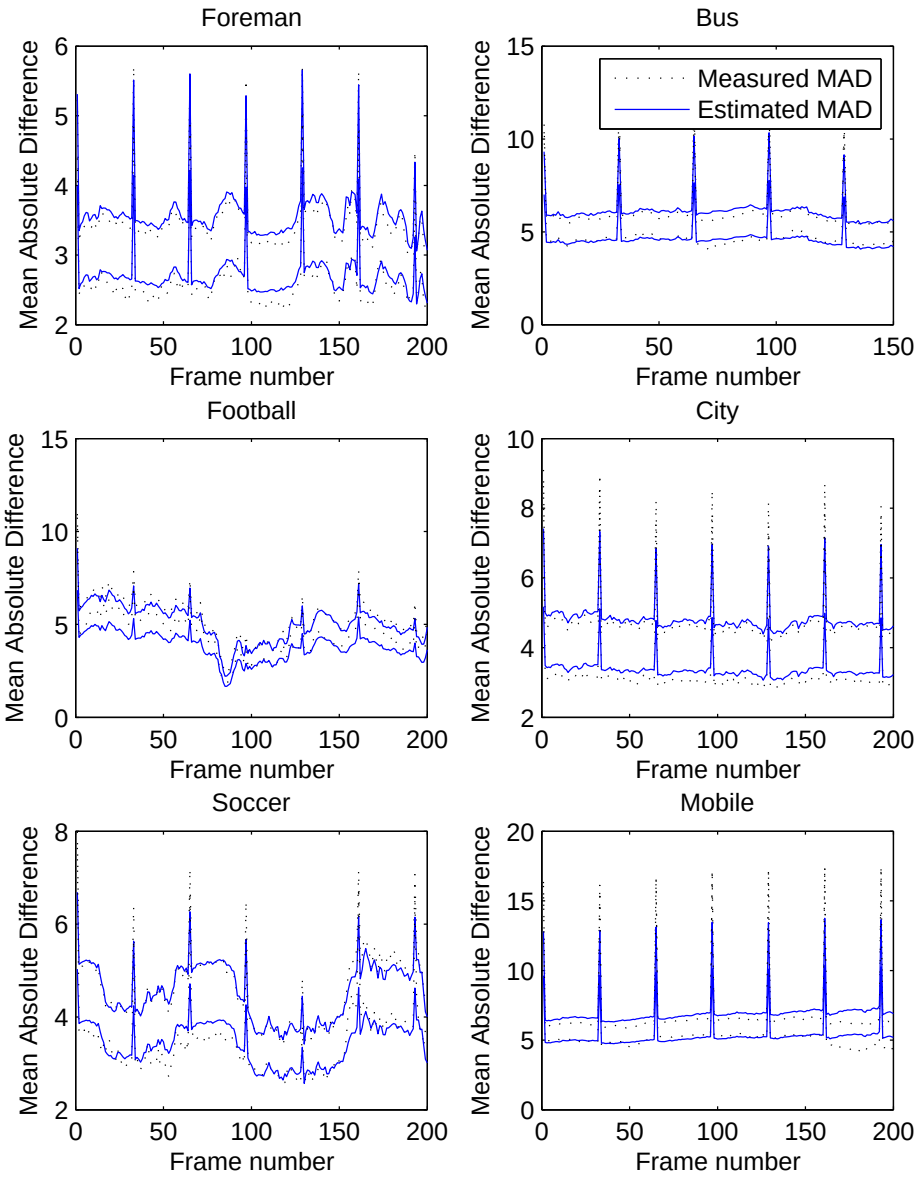
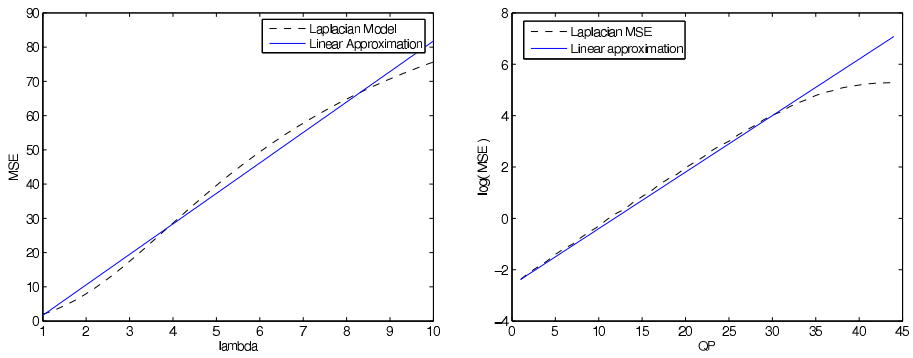


Fig. 4. Comparison between the predicted MAD (46) and true MAD values of two CGS enhancement layers encoded at QP = 32 and QP = 26. The prediction is performed using the base layer MAD encoded at QP = 38.



(a) (b)

Fig. 5. Relationship of the derived MSE with respect to the parameter λ and the quantization parameter QP. The figures show a linear relationship of the MSE with λ and an exponential relationship with the QP. The plots correpond to encoding 200 frames of the Foreman sequence at CIF resolution with an IDR period of 32 frames.

base-layer SVC stream, we propose the following PSNR expression:

$$D_b(\tilde{\lambda}, p_b) = b_1 \log_{10}\left((\tilde{\lambda}_b)^r + 1\right) \cdot p_b + b_2, \tag{48}$$

where r is a value that depends on the frame type, b_1 and b_2 are sequence-dependent model parameters typically valued at 0.52 and 47, respectively. The parameters b_1 and b_2 can be refined for each sequence during encoding. The value of r depends on the frame type, such that $r = 1$ for Inter-coded frames and $r \approx \frac{5}{6}$ for Intra-coded frames.

For CGS scalability in SVC, the term $\tilde{\lambda}(p_b)$ in (48) is simply replaced by the estimate of the quality refined prediction error $\tilde{\lambda}(p_e)$, thus resulting in the following CGS PSNR model:

$$D_e(\tilde{\lambda}, p_e) = b_1 \log_{10}\left((\tilde{\lambda}_e)^r + 1\right) \cdot p_e + b_2, \tag{49}$$

where $\tilde{\lambda}_e = \tilde{\lambda}_b 2^{a(p_e - p_b)}$, and b_1 and b_2 are the same constants used in the base layer model.

C. Bit-Rate Model

In this section, we present the bit-rate model for base and CGS enhancement layer SVC coded video frames. The presented model is based on the work proposed in [1] which we extended to incorporate CGS scalability. The H.264/AVC compatible base layer in SVC can be expressed as follows:

$$\begin{aligned} R_b(\tilde{\lambda}, q_b) &= c_1(\tilde{\lambda}_b)^r / q_b, \\ &= c_2(\tilde{\lambda}_b)^r 2^{-p_b/6} \end{aligned} \tag{50}$$

where c_1 is a model parameter, q_b is the quantization step size at the base layer, and r is the same power factor described in (48). Note that instead of the lookup table conversion from quantization parameter to quantization step described in the H.264/AVC standard, we approximate the conversion as follows:

$$q = 0.625 \cdot 2^{p/6}.$$

We extend the model in (50) to incorporate CGS packets which only contain refinements on the quantization of residual texture information [3]. Therefore, we express the enhancement layer bit-rate as follows:

$$R_e(\tilde{\lambda}, p_b) = c_3(\tilde{\lambda}_e)^r 2^{-p_e/6}. \tag{51}$$

VI. Performance Evaluations

In this section we demonstrate the accuracy of the proposed models. For our comparisons, we used the JSVM-9-8 reference software [19] to encode six reference video sequences in CIF resolution: Foreman, Bus, Football, City, Soccer, and Mobile. Each sequence is composed of 200 frames (Bus is composed of 150 frames) encoded in SVC, with an H.264/AVC compatible base layer, a GOP size of one, an IDR period of 32 frames, a frame rate of 30 fps. The bitstreams are comprised of a base layer encoded with QP = 38, and two CGS enhancement layers with QPs of 32 and 26, respectively. Note that by GOP we mean the group of pictures in a temporal decomposition block of SVC. By setting the GOP size to one, we eliminate the hierarchical prediction structure and limit our tests to a single temporal level comprised of I-frames and P-frames only.

A. Analytical and Empirical Models

The performance of the analytical base layer models can be seen in Figs. 6 and 7. The enhancement layer models are demonstrated in Figs. 8 and 9. The empirical models are demonstrated in Fig. 10. The performance is shown for the Foreman, Bus, Soccer, and Mobile sequences. The linear scaling parameters for these models are derived using the first 10 frames of each of the video sequences.

B. Comparison with Existing Models

We compare the performance of our proposed models with that of existing models described in Section II. Of these, we are interested in the real-time rate and distortion models proposed in [1], [8], [9].

These models illustrate three directions to rate-distortion modeling. In [1], the SATD of the transform coefficients as well as the quantization step size are the two main parameters used in the modeling. We will refer to these models by SATD. In [9], the transform coefficients are assumed to be Cauchy distributed. We will refer to these models as Cauchy. In [8], the transform coefficients are assumed to follow a Laplacian distribution as shown in (7). The distortion is derived as the error in quantizing a Laplacian distributed source using a scalar quantizer which parallels our derived model. The bit-rate is estimated as a weighted sum of the ℓ_0 and ℓ_1 norms of the quantized transform coefficients given by $N_{\rm nz}$ and $E_{\rm QTC}$, respectively. Our derivations have shown that the entropy of a quantized Laplacian source derived in (35) includes the $N_{\rm nz}$ and $E_{\rm QTC}$ terms as shown below:

$$\begin{aligned} H(q_w) \;=\; & -(1-e^{-(1-f)q_w/\hat{\lambda}})\log_2(1-e^{-(1-f)q_w/\hat{\lambda}}) \\ & -e^{-(1-f)q_w/\hat{\lambda}}\left[\log_2(1-e^{-q_w/\hat{\lambda}})-1\right] \\ & +N_{\rm nz}\tfrac{f}{\ln 2} + E_{\rm QTC}(1-e^{-q_w/\hat{\lambda}}). \end{aligned} \tag{52}$$

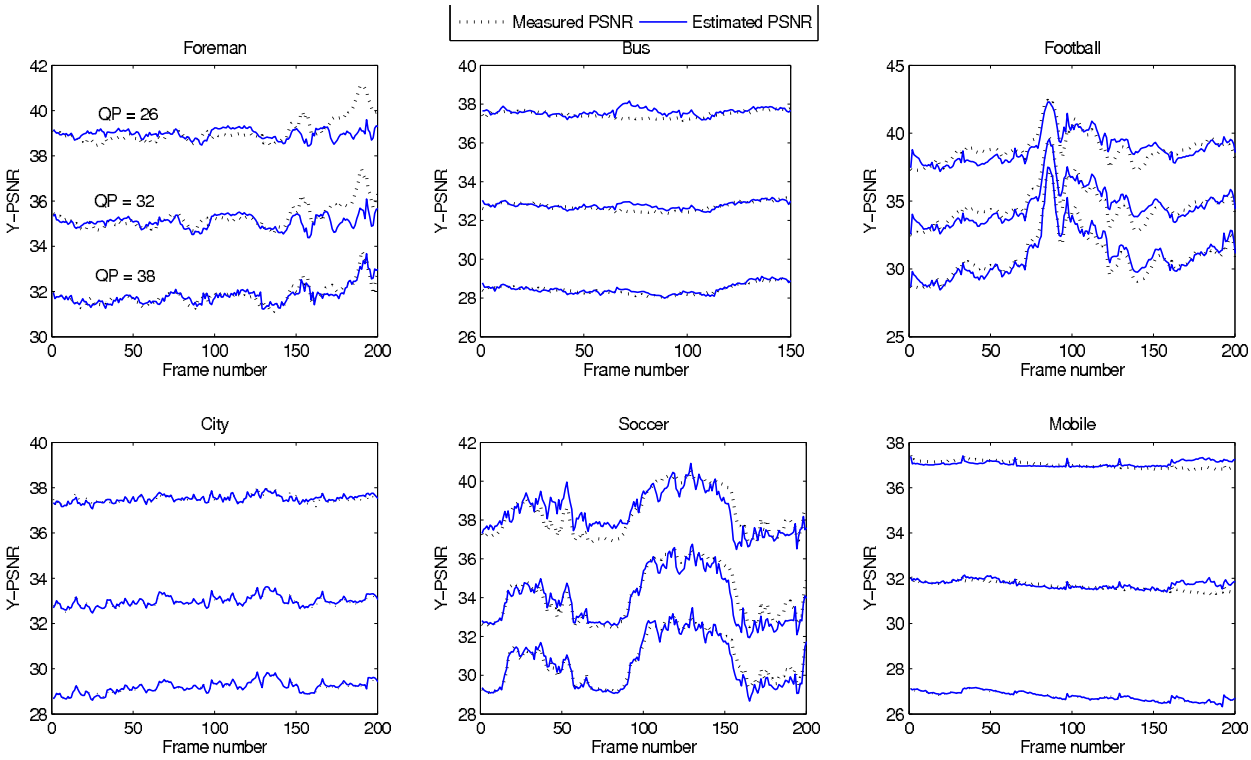


Fig. 6. Performance of the derived distortion model (31) for six reference video sequences encoded at QPs of 38, 32, and 26.

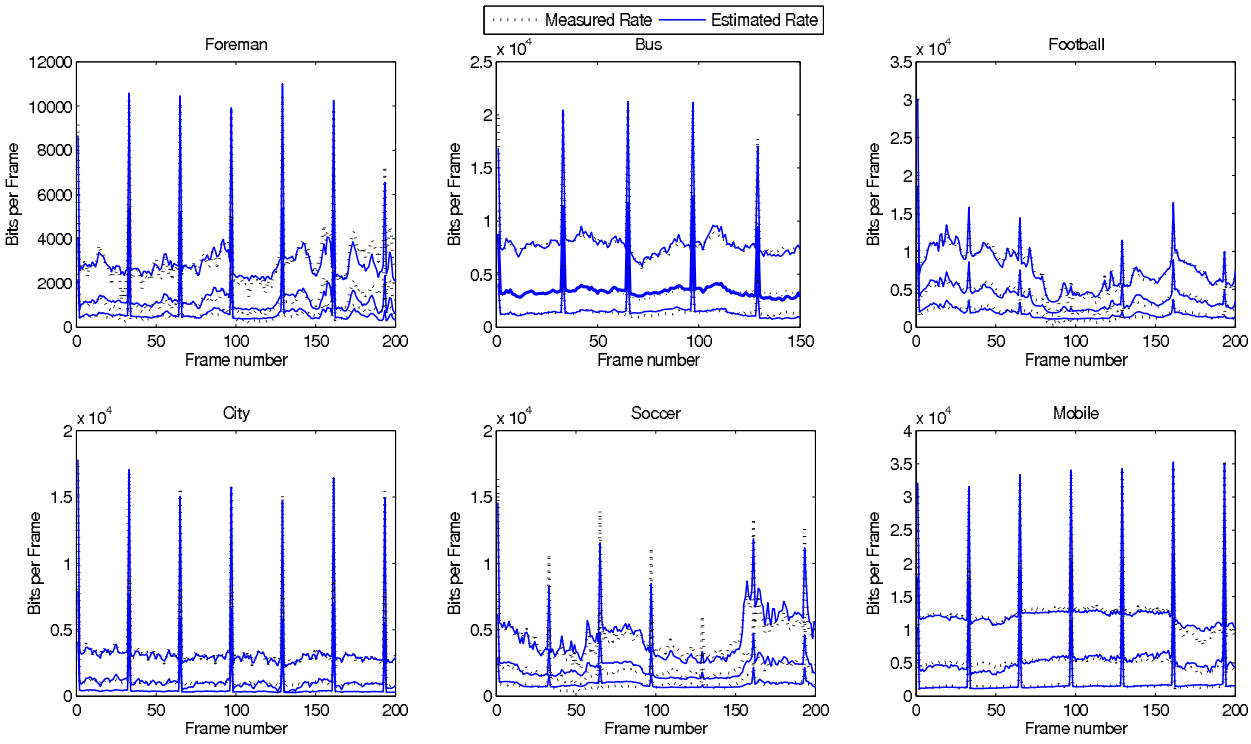


Fig. 7. Performance of the derived bit-rate model (36) for six reference video sequences encoded at QPs of 38, 32, and 26.

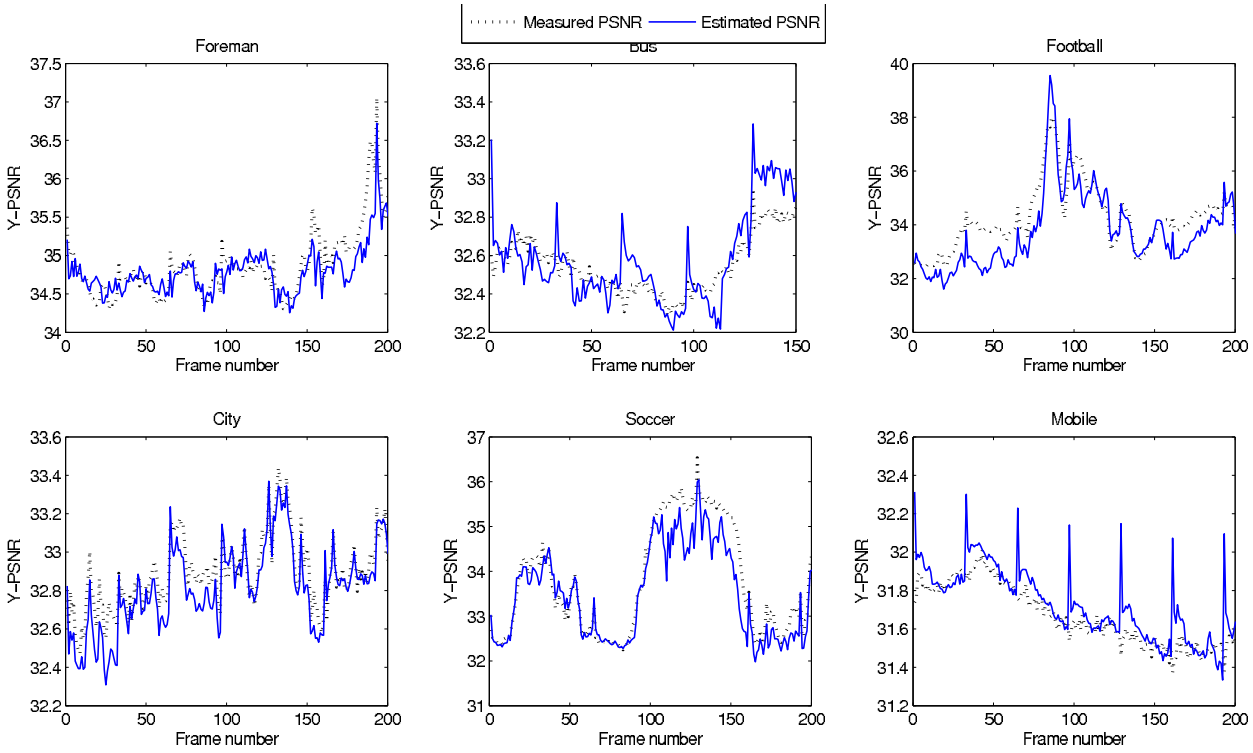


Fig. 8. Performance of the derived scalable distortion model (40) for the enhancement layer distortion of six reference video sequences with a base layer encoded at a QP = 38 and CGS enhancement layer encoded at a QP = 32.

Therefore, the models presented in [8] are simplified versions of our derived models where the simplification lies in the rate estimation and in assuming a fixed rounding operator $f = 1/6$.

Table II shows the root mean squared error (RMSE) of the rate and distortion estimation between the actual video bit-rate and PSNR, and the estimated bit-rate and PSNR given by our proposed model, the SATD model, and the Cauchy model. Table III lists the means and standard deviations of the bit-rate and PSNR of each of the video sequences included in Table II.

The results shown in Table II emphasize the accuracy of our proposed models especially with distortion estimation. The minimum RMSE value for each sequence is shown in bold font. It can be seen from the table that our proposed bit-rate model has a similar, if not slightly better performance, than the SATD bit-rate model. Meanwhile, our proposed distortion model outperforms SATD in all cases. It is worth mentioning that one advantage of our proposed models is that they are based on pixel-domain statistics whereas the SATD models are based on transform domain statistics which require additional computational complexity.

Concerning the Cauchy distribution based models, we have found several derivation errors made in [9]. These include finding the closed form solutions of Cauchy based entropy and the Laplacian based entropy shown in that

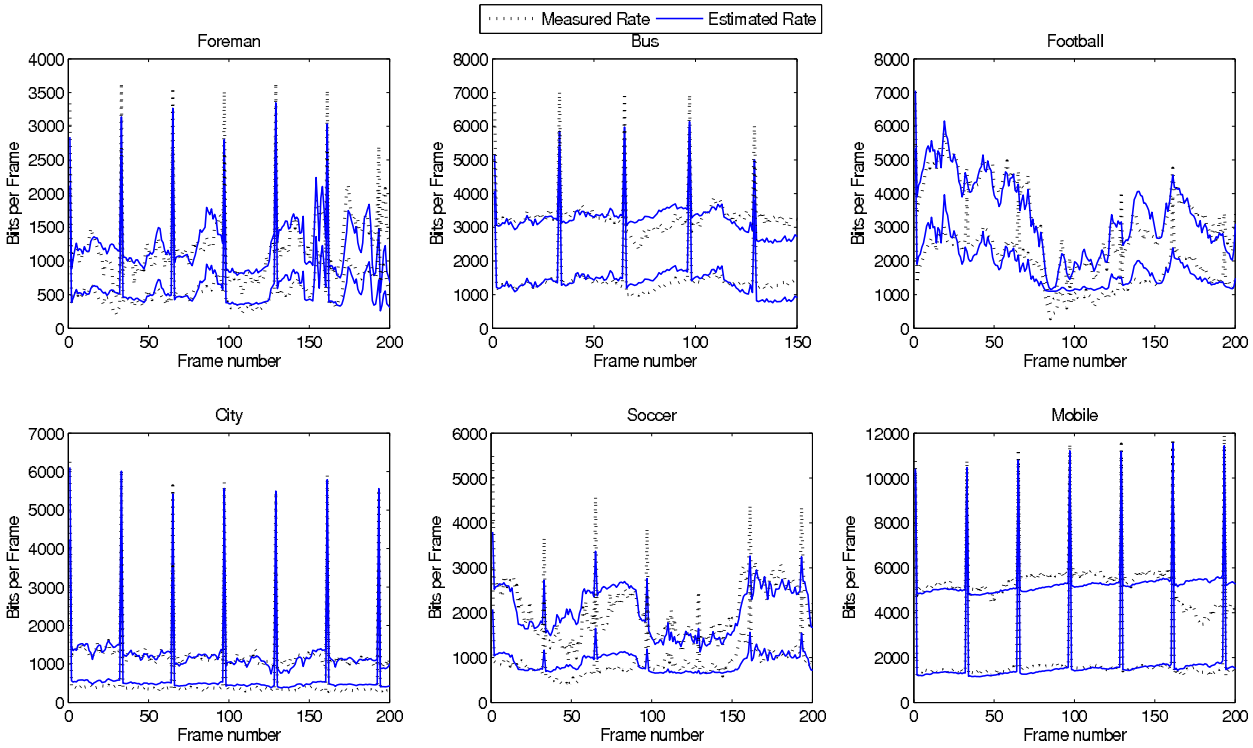


Fig. 9. Performance of the derived scalable bit-rate model (42) for six reference video sequences with the base layer encoded at QP = 38 and CGS enhancement layer encoded at QP = 32.

paper. We provide the corrected distortion and entropy expressions below:

$$\begin{aligned} D(\mu, q_w) &= \mu \tfrac{q_w}{\pi} - \tfrac{2\mu^2}{\pi} \tan^{-1}(\tfrac{q_w}{2\mu}) \\ &\quad +2 \sum_{i=1}^{\infty} \Big[\mu \tfrac{q_w}{\pi} - i\mu \tfrac{q_w}{\pi} \ln \left(\tfrac{\mu^2 + (i+1-f)^2 q_w^2}{\mu^2 + (i-f)^2 q_w^2} \right) \\ &\quad\quad - \tfrac{\mu^2 - i^2 q_w^2}{\pi} \left(\tan^{-1}((i+1-f)\tfrac{q_w}{\mu}) - \tan^{-1}((i-f)\tfrac{q_w}{\mu}) \right) \Big] \end{aligned} \tag{53}$$

$$\begin{aligned} H(\mu, q_w) &= -\tfrac{2}{\pi} \tan^{-1}((1-f)\tfrac{q_w}{\mu}) \log_2 \left(\tfrac{2}{\pi} \tan^{-1}((1-f) q_w/\mu) \right) \\ &\quad -\tfrac{2}{\pi} \sum_{i=1}^{\infty} \Big[\left(\tan^{-1}((i+1-f)\tfrac{q_w}{\mu}) - \tan^{-1}((i-f)\tfrac{q_w}{\mu}) \right) \\ &\quad\quad \times \log_2 \left(\tan^{-1}((i+1-f)\tfrac{q_w}{\mu}) - \tan^{-1}((i-f)\tfrac{q_w}{\mu}) \right) \Big] \end{aligned} \tag{54}$$

These errors might explain the reported improvement over the Laplacian based. However, our experiments have shown that the Cauchy-based models still failed to outperform our proposed models even after correcting the Cauchy derivations. Moreover, the simplified Cauchy models presented in [9] are an exact match to the SATD models presented in [1].

Finally, we are not aware of any works in the literature that attempt to model the rate and distortion behavior of CGS enhancement layers. Therefore, we have no benchmark to compare against.

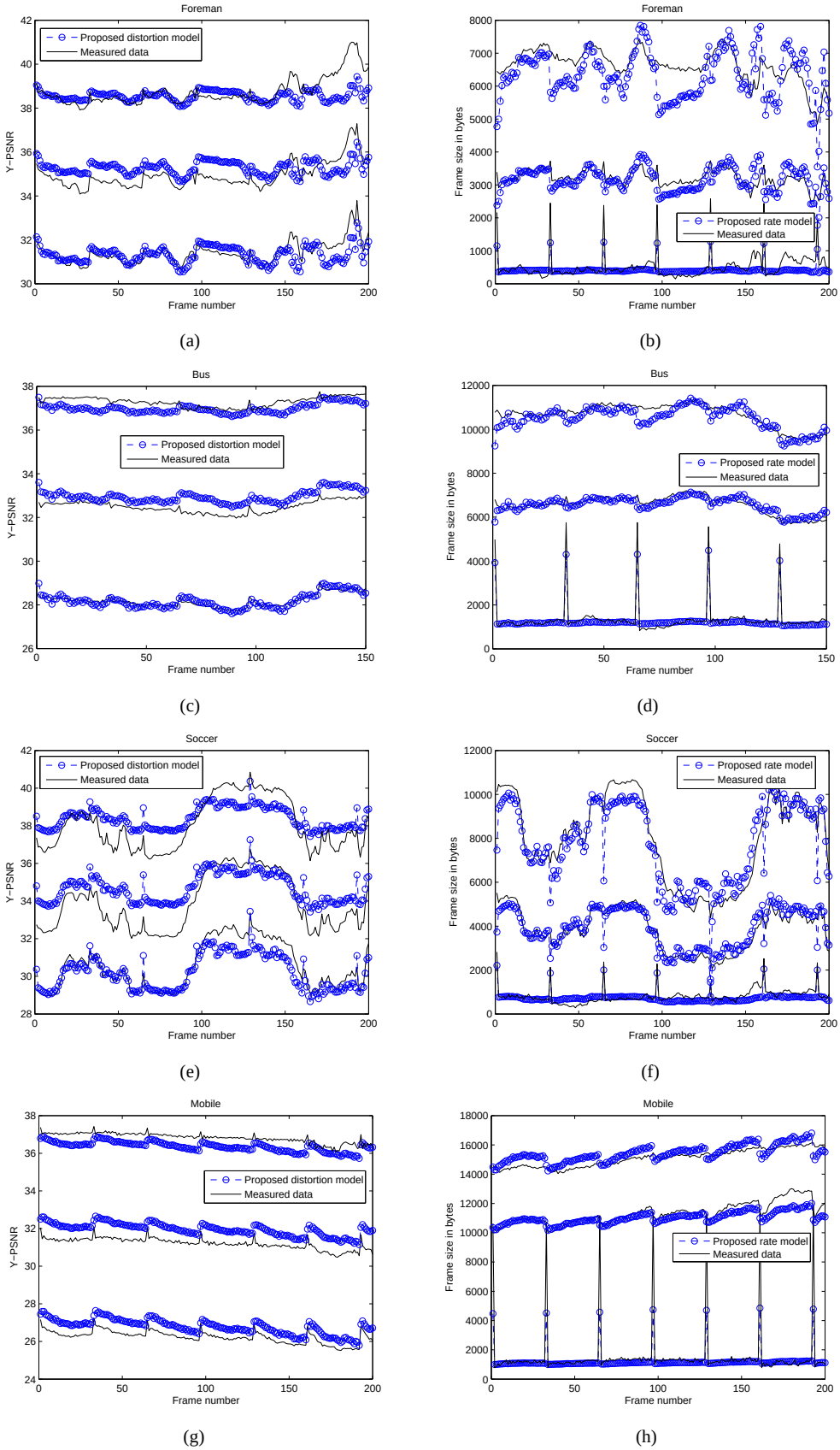


Fig. 10. Comparison between the simplified empirical distortion (left column) and bit-rate (right column) models and the actual data.

TABLE II
COMPARISON OF BIT-RATE (SIZE IN BITS) AND DISTORTION (PSNR) ESTIMATION IN TERMS OF RMSE

Sequence	QP	Bit-Rate				Distortion			
		Proposed	SATD	Cauchy	Tu et al. [8]	Proposed	SATD	Cauchy	Tu et al. [8]
Foreman	38	**193.8**	238.5	280.5	381.2	**0.2214**	0.2429	0.3695	0.2246
	32	336.7	**329.1**	755.9	735.7	**0.5162**	0.6972	0.5614	0.5240
	26	**415.6**	647.5	507.3	1044.3	**0.5783**	1.1108	0.6515	0.5987
Bus	38	286.1	**103.8**	815.1	737.5	**0.1165**	0.3351	0.2481	0.1310
	32	322.1	**200.1**	469.2	1077.1	**0.1871**	0.6336	0.2368	0.2203
	26	**393.4**	1130.4	601.2	1456.1	**0.2761**	0.4629	0.3786	0.3406
Football	38	438.8	**414.4**	763.3	489.9	**0.5228**	1.3883	2.4375	0.5240
	32	723.6	1023.4	827.4	**662.6**	**0.6222**	0.6368	2.4425	0.6274
	26	870.6	1355.9	1191.1	**835.7**	**0.6779**	0.6881	1.9159	0.6824
City	38	**48.5**	64.7	615.0	342.9	**0.0617**	0.3386	0.4777	0.0904
	32	359.3	**121.7**	294.3	977.52	**0.0768**	0.4324	0.3572	0.1404
	26	**148.3**	192.8	277.9	834.99	**0.1075**	0.5000	0.2623	0.2289
Soccer	38	**255.5**	292.2	320.1	345.8	**0.3191**	0.5934	1.1376	0.3239
	32	**517.3**	2159.2	835.0	724	**0.4864**	3.3382	1.9868	0.4932
	26	**705.4**	902.7	763.4	1171.9	**0.6547**	0.9335	1.6076	0.6708
Mobile	38	175.1	**128**	224.2	757	**0.0464**	0.1676	1.0436	0.1349
	32	711	**367**	510.1	2119.9	**0.1907**	0.5492	0.2605	0.2547
	26	**645.1**	13274	1689.3	2267.1	**0.1925**	6.0610	0.2210	0.2776

TABLE III
MEAN AND STANDARD DEVIATION OF THE BIT-RATE (SIZE IN BITS) AND DISTORTION (PSNR) OF THE TEST VIDEO SEQUENCES

Sequence	Foreman			Bus			Football			City			Soccer			Mobile		
QP	38	32	26	38	32	26	38	32	26	38	32	26	38	32	26	38	32	26
Bit-Rate ($\times 10^3$ bits)																		
Mean	0.6	1.3	3.1	1.5	3.6	8.1	1.7	3.8	7.6	0.5	1.2	3.5	0.9	2	4.5	1.8	5.5	13
Std. Dev.	0.4	0.8	1.4	0.8	1.4	2.3	0.7	1.6	2.8	0.6	1.4	2.4	0.4	0.9	1.9	1.8	2.9	4.1
PSNR (dB)																		
Mean	32	35	39	28	33	37	31	35	39	29	33	37	31	34	38	27	32	37
Std. Dev.	0.4	0.5	0.6	0.3	0.2	0.2	1.8	1.4	1	0.2	0.2	0.1	1.2	1.3	1.1	0.2	0.2	0.1

VII. CONCLUSION

In this paper, we derived single layer and scalable video rate and distortion models based on an assumption that the transform coefficients follow a Laplacian distribution. The models are customized to capture the behavior of a scalable video bitstream encoded using the CGS feature of SVC. Moreover, the models compensate for errors in the distribution assumptions by linearly scaling the Laplacian distribution parameter λ. Furthermore, we present simplified approximations of the derived models that allow for a run-time calculation of sequence dependent model constants. Our models use the mean absolute difference (MAD) of the residual signal extracted early in the encoding process and the quantization parameter (QP) values to estimate the residual MAD, rate, and distortion of a coded video frame at any QP value and for either base-layer and CGS layer packets. We have also shown performance evaluations that demonstrate that our proposed models accurately estimate the aforementioned metrics.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their invaluable comments and constructive feedback. Their input has been extremely appreciated and has helped in improving the quality and presentation of this work.

REFERENCES

[1] D.-K. Kwon, M.-Y. Shen, and C. C. J. Kuo, "Rate Control for H.264 Video With Enhanced Rate and Distortion Models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 5, pp. 517–529, May 2007.

[2] Z. Li, F. Pan, K. P. Lim, X. Lin, and S. Rahardja, "Adaptive rate control for H.264," in *Proceedings of IEEE International Conference on Image Processing*, 2004, pp. 745–748.

[3] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, September 2007.

[4] H. Mansour, P. Nasiopoulos, and V. Krishnamurthy, "Real-Time Joint Rate and Protection Allocation for Multi-User Scalable Video Streaming," in *Proceedings of IEEE Personal, Indoor, and Mobile Radio Communications (PIMRC)*, September 2008, pp. 1–5.

[5] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, "Analysis of Video Transmission over Lossy Channels," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1012–1032, June 2000.

[6] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, November 1998.

[7] X. Zhu and B. Girod, "Analysis of Multi-User Congestion Control For Video Streaming Over Wireless Networks," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, July 2006.

[8] Y.-K. Tu, J.-F. Yang, and M.-T. Sun, "Rate-Distortion Modeling for Efficient H.264/AVC Encoding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 5, pp. 530–543, May 2007.

[9] N. Kamaci, Y. Altunbasak, and R. Mersereau, "Frame bit allocation for the H.264/AVC video coder via Cauchy-density-based rate and distortion models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 8, pp. 994–1006, Aug. 2005.

[10] M. Dai, D. Loguinov, and H. Radha, "Rate-distortion modeling of scalable video coders," in *Proceedings of the IEEE International Conference on Image Processing ICIP '04*, vol. 2, October 2004, pp. 1093–1096.

[11] Z. He and S. Mitra, "Optimum bit allocation and accurate rate control for video coding via ρ-domain source modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 10, pp. 840–849, October 2002.

[12] *Coding of audio-visual objects Part 2: Visual*, ISO/IEC 14492-2 (MPEG-4 Visual), ISO/IEC JTC 1, May 2004.

[13] H. S. Malvar, A. Hallapuro, M. Karczewicz, and L. Kerofsky, "Low-Complexity Transform and Quantization in H.264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 598–603, July 2003.

[14] *Advanced Video Coding for Generic Audiovisual Services*, ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), ITU-T and ISO/IEC JTC 1, Version 6: June 2006.

[15] Y. Su, J. Xin, A. Vetro, and H. Sun, "Efficient MPEG-2 to H.264/AVC intra transcoding in transform-domain," in *Proceedings of the IEEE International Symposium on Circuits and Systems, ISCAS 2005*, vol. 2, May 2005, pp. 1234–1237.

[16] F. J. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[17] G. Sullivan, T. Wiegand, and K.-P. Lim, "Joint Model Reference Encoding Methods and Decoding Concealment Methods," ISO/IEC JTC1/SC29/WG11 and ITU-T Q6/SG16 JVT-I049, September 2003.

[18] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Transactions on Circuits and Systems for Video Technology,*, vol. 13, no. 7, pp. 688 – 703, July 2003.

[19] ISO/IEC JTC 1/SC 29/WG 11 N8964, "JSVM-10 software," 2007. [Online]. Available: http://wg11.sc29.org/mpeg/docs/_listwg11_80.htm