Joint Baseband Signal Quantization and Transform Coding for High Dynamic Range Video

Chau-Wai Wong, *Member, IEEE,* Guan-Ming Su, *Senior Member, IEEE,* and Min Wu, *Fellow, IEEE*

***Abstract*—Digitally acquired high dynamic range (HDR) video baseband signal can take 10 to 12 bits per color channel. It is of economical importance to be able to reuse the legacy 8 or 10-bit video codecs to efficiently compress the HDR video. Linear or nonlinear mapping on the intensity can be applied to the baseband signal to reduce the dynamic range before the signal is sent to the codec, and we refer to this range reduction step as the baseband quantization. We show analytically and verify using test sequences that the existence of the baseband quantizer lowers the coding efficiency. Experiment shows that as the baseband quantizer strengthened by 1.6 bits, the drop of PSNR at high bitrate is up to 1.60 dB. Our result suggests that, in order to achieve high coding efficiency, video information reduction in terms of quantization error should be incurred in the video codec instead of on the baseband signal. (Version: 22:46 EST, Tuesday 8th March, 2016)**

***Index Terms*—Quantization, High Dynamic Range (HDR), Bitdepth, Transform Coding, HEVC/H.265**

I. Introduction

The need for more vivid digital videos relies on two main factors: more pixels, and better pixels [1], [2]. The latter is more important than the former when nowadays the resolution goes beyond the high definition. At the signal level, the need for better pixels means adopting a wide color gamut (WCG), and using high dynamic range (HDR) to represent all colors with small quantization errors [3]–[7].

One efficient color coding standard that keeps the visibility of quantization artifacts to a uniformly small level is the perceptual quantizer (PQ) [8], [9], but it still takes 12 bits to represent all luminance levels. It is of economical importance to be able to reuse the legacy 8 or 10-bit video codecs such as H.264/AVC [10] and H.265/HEVC [11] to efficiently compress HDR videos. Linear or nonlinear mapping on the intensity can be applied to the baseband signal to reduce the dynamic range before the signal is sent to the codec, and we refer to this range reduction step as the baseband quantization. Even if a codec supports the dynamic range of a video, range reduction can also be motivated by the needs of i) saving the running time of the codec via computing numbers in a smaller range, ii) handling the event of instantaneous bandwidth shortage as a coding feature provided in VC-1 [12]–[14], or iii) removing color precision that cannot be displayed by old screens.

Hence, it is important to ask whether reducing bitdepth for baseband signal is bad for coding efficiency measured in HDR. Practitioners would say “yes”, but if one starts to tackle this question formally, the answer is not immediately clear as the change of the rate-distortion (RD) performance is non-trivial: reducing the bitdepth for baseband signal while maintaining the compression strength of the codec will lead to a smaller size of encoded bitstream and a larger error measured in HDR.

We approach this problem by establishing the relationship between the strength of the baseband quantizer and the coding efficiency measured in (peak) signal-to-noise ratio [(P)SNR]. It is beneficial to first model the problem of quantifying the error in reconstructed images [15] as the problem of quantifying the error in reconstructed residues. We then examine the error of a single quantizer, and arrive at Lemma 2 that serving as a primitive to facilitate the joint analysis on the effects of baseband and codec quantizers with a linear transform.

The paper is organized as follows. In Section II, we simplify the practical HDR video coding pipeline into a theoretically tractable model before diving into the main derivation in Section III-A. Simulation results are presented in Section III-B to validate the derivation, and experimental results on videos are presented in Section IV to confirm the theoretical explanation.

II. HDR Video Coding Pipeline Modeling

A. Quantifying Frame Error by Residue Error

Block diagram shown in Fig. 1 (a) models the video coding pipeline with the effect of baseband signal quantization. The input to the pipeline is the HDR frame at time index t, $\mathbf{I}_t^{\mathrm{HDR}}$, with L pixels. The immediate input to the video codec $\mathbf{I}_t$ and final reconstructed output $\hat{\mathbf{I}}_t^{\mathrm{HDR}}$ are limited by the precision of the finite bits container, so pixels take values on the set $q_1\mathbb{Z} = \{nq_1 | n \in \mathbb{Z}\}$. The immediate output pixels from the codec take integer values due to the rounding operation at the final stage of the codec, and the integer-valued vector $\hat{\mathbf{I}}_{t-1}$ is used by intra- and inter-predictors collectively modeled as $\mathrm{pred}(\cdot)$. For simplicity, we define the quantizer function $\mathcal{Q}_i(x) = \mathrm{iQ}_i\left(\mathrm{Q}_i(x)\right)$ before stating the following lemma:

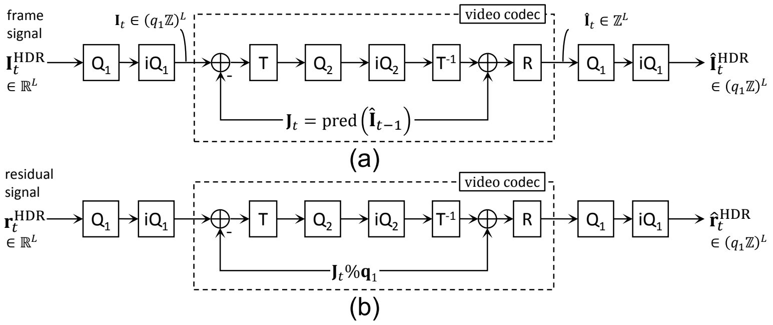
Lemma 1 (frame error by residue error)**.** *The problem of quantifying the error of predictively coded video frames can be reduced approximately to quantifying the error of non-predictively coded residues.*

The sketch of the proof is described here, the complete proof is left to Appendix A. We first establish the equivalence of the two pipelines of Figs. 1 (a) and (b), where $\mathbf{r}_t^{\mathrm{HDR}} \overset{\mathrm{def}}{=} \mathbf{I}_t^{\mathrm{HDR}} - \mathcal{Q}_1(\mathbf{J}_t)$. Then, the assumption that quantization step of Q_1 is

C.-W. Wong and Min Wu are with the Department of Electrical and Computer Engineering, and the Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA. This work was initiated when C.-W. Wong was a research intern at Dolby Laboratories in 2014. E-mail: (cwwong, minwu)@umd.edu.

G.-M. Su is with Dolby Laboratories, Sunnyvale, CA 94085, USA. E-mail: guanmingsu@ieee.org.

Manuscript received March 8th, 2016.



frame signal
$\mathbf{I}_t^{\text{HDR}} \in \mathbb{R}^L$
$\mathbf{I}_t \in (q_1\mathbb{Z})^L$
video codec
$\hat{\mathbf{I}}_t \in \mathbb{Z}^L$
Q1
iQ1
T
Q2
iQ2
T-1
R
$\hat{\mathbf{I}}_t^{\text{HDR}} \in (q_1\mathbb{Z})^L$
$\mathbf{J}_t = \text{pred}(\hat{\mathbf{I}}_{t-1})$
(a)
residual signal
$\mathbf{r}_t^{\text{HDR}} \in \mathbb{R}^L$
$\hat{\mathbf{r}}_t^{\text{HDR}} \in (q_1\mathbb{Z})^L$
$\mathbf{J}_t\%\mathbf{q}_1$
(b)

Fig. 1: (a) Block diagram for the video coding process with the effect of baseband signal quantization, and (b) equivalent diagram of (a). Block R is the rounding to the nearest integer operation, round(x). $\text{Q}_i(x) \overset{\text{def}}{=} \text{round}\,(x/q_i)\,, \text{iQ}_i(x) \overset{\text{def}}{=} q_i \cdot x,\ \ i = 1, 2$ are quantization and dequantization, respectively. All operations are applied separately to each entry of x when x is a vector.

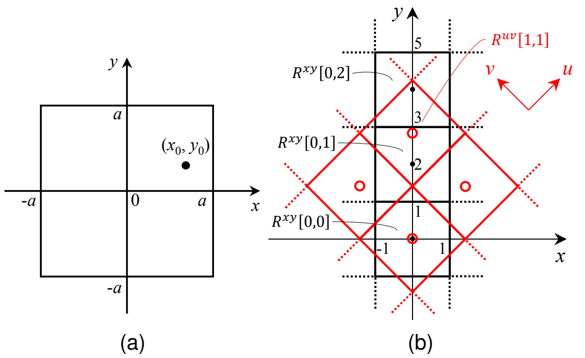


Fig. 2: Illustration for MSE calculation for the cases that (a) any point (x, y) located within the $2a$-by-$2a$ square is quantized to the reconstruction centriod (x_0, y_0) that may or may not located within the square, and (b) a quantization in xy-plane is followed by a transform, a quantization in uv-plane, inverse transform, and a quantization in xy-plane.

much smaller comparing to the range of $\mathbf{r}_t^{\text{HDR}}$ allows us to remove the predictive branch of Fig. 1 (b), and to declare the non-predictive coding branch of Fig. 1 (b) is approximately equivalent to the original pipeline of Fig. 1 (a).

B. Quantization Error for a Hypercube

Assume that the reconstruction centroid for a squared region of edge length $2a$ centered at $(0, 0)$[1] as shown in Fig. 2 (a) is located at $(x_0, y_0) \in \mathbb{R}^2$, not limited to be within the region. We further assume that the point (X, Y) is uniformly distributed over the square, namely, the joint distribution $f_{X,Y}(x, y) = \frac{1}{4a^2},\ (x, y) \in [-a, a]^2$. The mean-squared error (MSE) for the random vector (X, Y) quantized to/reconstructed at (x_0, y_0) is

$$\begin{aligned} \text{MSE} &= \mathbb{E}\left[\|(X, Y) - (x_0, y_0)\|^2\right] \\ &= \int_{-a}^{a}\int_{-a}^{a} \|(x, y) - (x_0, y_0)\|^2 f_{X,Y}(x, y)\, dx\, dy \\ &= \frac{1}{4a^2}\int_{-a}^{a} dy \int_{-a}^{a} (x - x_0)^2 + (y - y_0)^2\, dx \\ &= d^2 + \frac{2}{3}a^2 \end{aligned} \tag{1}$$

where $d^2 = x_0^2 + y_0^2$ is the squared Euclidean distance to the geometric center of the region, $(0, 0)$, and $\frac{2}{3}a^2$ is related to the strength of the quantizer. It is straight forward to extend the result to the N-dimensional (N-d) case shown as follows:

Lemma 2 (quantization error)**.** *The mean-squared error for a point that is uniformly distributed within an N-d hypercube with an arbitrarily positioned reconstruction centroid and edge length $2a$ is $d^2 + \frac{N}{3}a^2$, where d is the Euclidean distance from the centroid to the geometric center of the hypercube.*

This result agrees with two intuitive observations. First, as the reconstruction centroid departs from the geometric center, the quantization error increases. Second, as the quantizer strength quantified by the edge length $2a$ increases, the error increases.

III. Effect of Baseband Quantizer on Coding Efficiency

A. Error of Video Coding With Baseband Quantizer

Lemma 1 allows us in the following analysis to avoid dealing with the predictive coding loop, and merely to follow a scheme with transform coding and quantization blocks in series. In addition, the residue signal that can be more easily modeled in the probabilistic sense than the frame signal is used as the input. Lemma 2 converts the derivation of the reconstruction error of all possible points to that of just a few reconstruction centroids.

We again use an example with two axes as shown in Fig. 2 (b) to illustrate the idea behind, and all the derivations can be easily expanded to the N-d general case.

Assume the input residual signal is a data point (x, y) on the xy-plane with a joint probability distribution $f_{XY}(x, y)$. Transform by an orthogonal matrix $\mathbf{T}$ can be considered geometrically as a rotation of the coordinate system, namely,

$$(x, y) \overset{\mathbf{T}}{\mapsto} (u, v), \quad [u\ v]^T = \mathbf{T}\,[x\ y]^T \tag{2}$$

where we choose $\mathbf{T} = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right)$. In this example, $(1, 0) \overset{\mathbf{T}}{\mapsto} (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ and $(1, 1) \overset{\mathbf{T}}{\mapsto} (\sqrt{2}, 0)$.

Quantization is equivalent to cutting the plane into squares, as shown in Fig. 2 (b). We denote the point set containing

[1] Throughout this paper, column vector $[x_1\ x_2\ \cdots\ x_n]^T$ may be denoted as $(x_1, x_2, \cdots, x_n)$ for the purpose of compact presentation.

all points belonging to a quantized region in xy-plane with horizontal index i and vertical index j as $\boldsymbol{R}^{xy}[i,j]$, where indices $i,j \in \mathbb{Z}_M \overset{\text{def}}{=} \{-M,\cdots,0,\cdots,M\}$. Geometric centers in of $\boldsymbol{R}^{xy}[i,j]$ are denoted as "$\bullet$", and those of $\boldsymbol{R}^{uv}[i,j]$ are denoted as "$\circ$". In this example, $\boldsymbol{R}^{xy}[0,1]$ centered at $(0,2)$ and $\boldsymbol{R}^{xy}[0,2]$ centered at $(0,4)$ are both quantized to $\boldsymbol{R}^{uv}[1,1]$ by $\mathcal{Q}_2$, and finally quantized to $\boldsymbol{R}^{xy}[0,1]$ by $\mathcal{Q}_1$.

The overall error $D \overset{\text{def}}{=} \mathbb{E}[\|(x,y)-(\hat{x},\hat{y})\|^2]$ due to video coding with baseband quantizer can be calculated by averaging MSE over the $(2M+1)^2$ regions indexed by (i,j), namely,

$$D = \mathbb{E}\Big[\mathbb{E}\big[\|(x,y)-(\hat{x},\hat{y})\|^2|\boldsymbol{R}^{xy}[I,J]\big]\Big] \tag{3}$$

where the probability mass function is $p_{IJ}(i,j) = \int_{x,y\in \boldsymbol{R}^{xy}[i,j]} f_{XY}(x,y)\,dxdy$. For each region $\boldsymbol{R}^{xy}[i,j]$, the calculation of error is simplified by Lemma 2 [2], namely,

$$\mathbb{E}\big[\|(x,y)-(\hat{x},\hat{y})\|^2|\boldsymbol{R}^{xy}[i,j]\big] = \frac{2}{3}\left(\frac{q_1}{2}\right)^2 + d^2\{\boldsymbol{R}^{xy}[i,j]\} \tag{4}$$

where $d\{\boldsymbol{R}^{xy}[i,j]\}$ is the Euclidean distance from the reconstruction centroid to the geometric center of $\boldsymbol{R}^{xy}[i,j]$. The geometric center is by definition $\mathbf{m} = (iq_1, jq_1)$. Passing $\mathbf{m}$ through the whole pipeline shown in Fig. 1 (b) excluding the predictive branch (*aka* the main branch), one can obtain the reconstruction centroid:

$$\hat{\mathbf{m}} = \mathcal{Q}_1\Big(\mathbf{T}^{-1}\left\{\mathcal{Q}_2\left[\mathbf{T}\,\mathcal{Q}_1\left([iq_1\ jq_1]^T\right)\right]\right\}\Big) \tag{5a}$$

$$= q_1\ \text{round}\left[\frac{q_2}{q_1}\mathbf{T}^{-1}\ \text{round}\left(\frac{q_1}{q_2}\mathbf{T}\begin{bmatrix} i\\ j\end{bmatrix}\right)\right]. \tag{5b}$$

Substituting Eqn. (4) into Eqn. (3), we obtain:

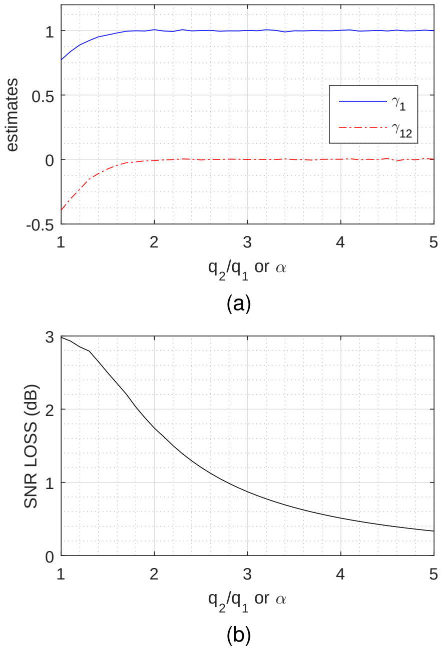
$$D = \frac{2}{3}\left(\frac{q_1}{2}\right)^2 + \mathbb{E}\big[d^2\{\boldsymbol{R}^{xy}[I,J]\}\big]. \tag{6}$$

Due to the space limitation, we leave the detailed derivation for $\mathbb{E}\left[d^2\{\boldsymbol{R}^{xy}[I,J]\}\right]$ to Appendix B. We present the final result of the derivation for the overall error D for scenarios that the baseband quantizer is finer than the codec quantizer (*i.e.*, $q_1 < q_2$) as follows:

$$D = \begin{cases} \frac{N}{12}\left(q_2^2 + 2q_1^2\right), & q_1 \le \frac{q_2}{2}, \quad (7)\\ \frac{N}{12}\left[q_2^2 + (1+\gamma_1)q_1^2 + 2\gamma_{12}q_2q_1\right], & q_1 > \frac{q_2}{2}, \quad (8)\end{cases}$$

where N is the length of input signal vectors, and estimates of γ_1 and γ_{12} are displayed in Fig. 3 (a).

It can be proved that, using the scheme of the main branch of Fig. 1 (b), the bitrate is solely controlled by the codec quantizer $\mathcal{Q}_2$. Hence, fixing q_2 and thus the bitrate, any increase in q_1 leads to a decrease in SNR and therefore in coding efficiency. In comparison, a change in q_2, which changes bitrate and SNR simultaneously, has no impact on the coding efficiency.[3]



(a)

(b)

Fig. 3: (a) Estimated γ_1 and γ_{12}, and (b) SNR LOSS as a function of q_2/q_1 or α.

Given the scenarios of interest that $q_1 < q_2$, we define $q_1 = \frac{q_2}{\alpha}$, $\alpha \ge 1$. The SNR loss with reference to an almost perfectly fine baseband quantizer, *i.e.*, $q_1 \to 0$, can be easily derived:

$$\text{SNR LOSS} = \begin{cases} 10\log_{10}\left(1+\frac{2}{\alpha^2}\right), & \alpha \ge 2, \quad (9)\\ 10\log_{10}\left(1+\frac{1+\gamma_1}{\alpha^2}+\frac{2\gamma_{12}}{\alpha}\right), & \alpha < 2. \quad (10)\end{cases}$$

The resulting SNR loss is shown in Fig. 3 (b).

To conclude, under the assumption of $q_1 < q_2$, i) the best case is $q_1 \ll q_2$ or $\alpha \to \infty$, and error is solely due to the codec quantizer and there is no reduction in SNR; and ii) the worst case is reached when $q_1 \nearrow q_2$ or $\alpha \searrow 1$, and a maximum of 3 dB SNR drop is incurred.

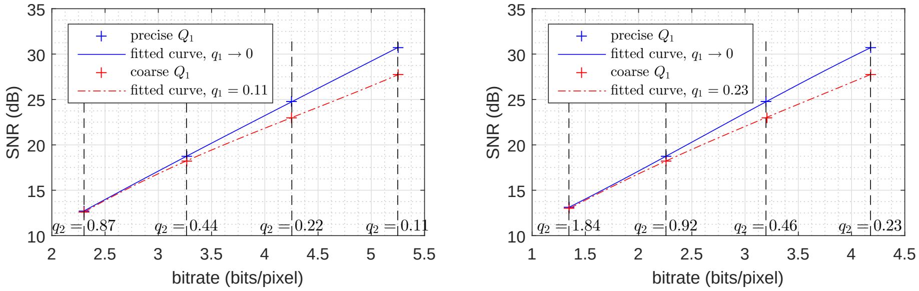
B. Simulation Results

We verify the theoretical result by simulating the change of SNR as a function of $\frac{q_1}{q_2}$. Specifically, assume a length-L Gaussian vector $(X_1, X_2, \cdots, X_L)$, with a fixed correlation of neighboring coordinates, *i.e.*, $\text{corr}(X_l, X_{l+1}) = \rho$, for $l = 1, \cdots, L-1$. In image/video coding scenarios, L usually takes value in $\{4^2, 8^2, 16^2\}$. In our simulation, realizations of random vectors are obtained by choosing disjoint segments from a realization of an AR(1) process.

We present two simulation cases, (a) for small blocks with low neighborhood correlation ($L = 4^2$, $\rho = 0.4$, $\sigma = \text{std}(X_l) = 1.0911$), and case (b) for large blocks with high correlation ($L = 16^2$, $\rho = 0.9$, $\sigma = 2.2942$). In both cases, we check the performance difference between the scenarios when the baseband quantizer $\mathcal{Q}_1$ is negligible, *i.e.*, $q_1 \to 0$, and not. When $\mathcal{Q}_1$ is not negligible, we set the quantizer to be reasonably coarse with respect to the spread of X_l, namely, $q_1 = \frac{\sigma}{10}$, and the corresponding q_2 for each bitrate value on RD curve from left to right are $q_1 \times \{8, 4, 2, 1\}$.

[2] Note that the uniform distribution assumption of Lemma 2 is valid within region $\boldsymbol{R}^{xy}[i,j]$ for the high bitrate coding scenario that we are interested in.

[3] Recall that the comparison of coding efficiency between two codecs is via the comparison of their empirical RD curves. A change in q_2 does lead to a move of the operation point in the bitrate-SNR plane, but both the starting and the ending locations reside on the same RD curve.



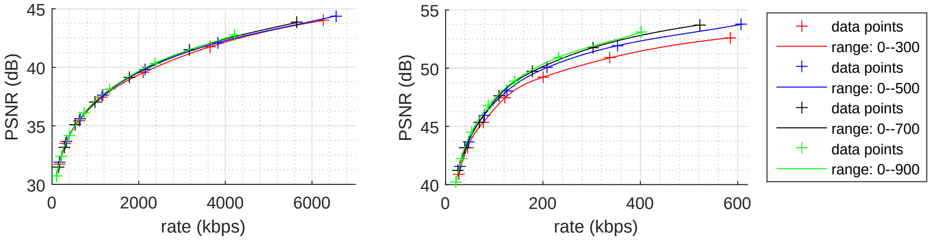
(a)

(b)

q_1/q_2	simulation case (a)	 case (b)	theory [Eqn. (9)]
1/8	−0.13	−0.16	−0.13
1/4	−0.51	−0.53	−0.51
1/2	−1.74	−1.75	−1.76
1/1	−3.00	−2.95	−2.98

(c)

Fig. 4: RD curves of simulated video data revealing the performance gap between the scenarios that the baseband quantizer $\mathcal{Q}_1$ is negligible (**solid blue**) and not negligible (**dotted red**). Case (a): small blocks with low neighborhood correlation ($L = 4^2$, $\rho = 0.4$), and case (b): large blocks with high correlation ($L = 16^2$, $\rho = 0.9$). (c) Simulated SNR drops for different ${}^{q_1}/_{q_2}$ ratios agree with the theoretical results [Eqn. (9)].



(a)

(b)

sequence	intra-coding PSNR gap (dB) at extreme BRs low	 high	inter-coding low	 high
BalloonFe.	0.13	0.83	0.19	0.83
Market	0.04	0.23	0.03	0.35
Typewriter	0.29	1.12	0.62	1.60

(c)

Fig. 5: RD curves for inter-coded videos with different strength levels of the baseband quantizer $\mathcal{Q}_1$: (a) Market and (b) Typewriter. The PSNR gaps at high bitrate are 0.35 and 1.60 dB. (c) Largest PSNR gaps at both high and low bitrates for intra- and inter-coded videos.

Simulation results are shown in Fig. 4. The solid red curves show the RD performances when the coarse quantizer $\mathcal{Q}_1$ is used; whereas the dotted blue curves show the performances when $\mathcal{Q}_1$ is negligible, which is to approximate the scenario that $\mathcal{Q}_1$ is absent. The RD performance drop with respect to the solid blue curves is consistent with the theoretical estimates shown in the table of Fig. 4 (c). As expected from our theoretical result, the above results are independent of block size L and neighborhood correlation ρ.

IV. Experimental Results on Videos

We now verify the theoretical results using standard test sequences. Test sequences stored in the 16-bit TIFF container are regarded as the reference/baseband signal. They were first linearly mapped to different dynamic ranges to mimic the effect of the baseband quantizer, the resulting videos were then encoded using HM 14.0 [16], and finally the quality in terms of PSNR and SSIM was measured in the 16-bit precision.

Detailed simulation conditions are as follows. The luma component of three test sequences BalloonFestival, Market, and Typewriter in BT.2020 color space [17] of size 1920×1080 are used. The operational bitdepth in the video codec is 10. Each video is encoded using two structures: the all I-frames structure for 17 frames (*aka* intra-coding), and the IBBB $\cdots$ structure for 64 frames (*aka* inter-coding). The codec quantizer takes 6 equally spaced quantization parameters to draw one piece of RD curve. Videos are baseband-quantized to the dynamic ranges $[0, 300], [0, 500], [0, 700]$, and $[0, 900]$ with effective bitdepth $8.2, 9.0, 9.5$, and 9.8 bits, respectively.

The experimental results from all sequences with both PSNR and SSIM measure reveal that the stronger the baseband quantizer is, the more penalty in coding efficiency is incurred. Due to the space limitation, we show the RD performance measured in PSNR for Market and Typewriter that are inter-coded in Figs. 5 (a) and (b). It can be read from the figures that the PSNR gaps between the green curve and the red curve at a high bitrate (the largest rate that 4 curves simultaneously cover) is 0.35 dB for Market and 1.60 dB for Typewriter. Table of Figs. 5 (c) reports the largest PSNR gaps at both high and low bitrates for intra- and inter-coded videos. It is observed that as the baseband quantizer strengthened by $1.6\,(= 9.8 - 8.2)$ bits, the drop of PSNR at a high bitrate is up to nearly 1.60 dB.

V. Conclusion and Discussion

In this work, we analyzed the video coding pipeline by explicitly considering the existence of the baseband quantizer. We arrived at the conclusion via theoretical proof and experiment that the baseband quantizer lowers the coding efficiency, whereas the codec quantizer does not affect the coding efficiency. Hence, video information reduction in terms of quantization error should be incurred in the video codec instead of on the baseband signal.

In a more practical scenario, nonlinear mapping is more often used than linear mapping for baseband signal range reduction when the bitdepth is insufficient. Although we have proved that quantizing the baseband signal uniformly leads to a penalty in coding efficiency measured in HDR, it is interesting to see whether quantizing the baseband signal non-uniformly can also lead to a penalty in coding efficiency.

References

[1] D. G. Brooks, “The art of better pixels,” *SMPTE Motion Imaging Journal*, vol. 124, no. 4, pp. 42–48, May 2015.

[2] S. Karlin, “New display technology aims to preserve realistic colors and contrasts,” *IEEE Spectrum*, Mar. 2014.

[3] G.-M. Su, Q. Chen, H. Koepfer, and S. Qu, “Joint base layer and enhancement layer quantizer adaptation in EDR video coding,” US Patent 9,219,916 B2, 2015.

[4] T. Lu, F. Pu, P. Yin, T. Chen, and W. Husak, “Implication of high dynamic range and wide color gamut content distribution,” in *Proc. SPIE, Applications of Digital Image Processing XXXVIII*, San Diego, CA, Aug. 2015, p. 95990B.

[5] P. Hanhart, M. Rerabek, and T. Ebrahimi, “Towards high dynamic range extensions of HEVC: subjective evaluation of potential coding technologies,” in *Proc. SPIE, Applications of Digital Image Processing XXXVIII*, San Diego, CA, Aug. 2015, p. 95990G.

[6] A. Luthra, E. Franois, and W. Husak, “Draft call for evidence (CfE) for HDR and WCG video coding,” ISO/IEC JTC1/SC29/WG11 MPEG, Strasbourg, France, Tech. Rep. N15028, Oct. 2014.

[7] “Test results of call for evidence (CfE) for HDR and WCG video coding,” ISO/IEC JTC1/SC29/WG11 MPEG, Warsaw, Poland, Tech. Rep. N15350, Jun. 2015.

[8] *ST 2084:2014, High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays*, SMPTE Std., Aug. 2014.

[9] S. Miller, M. Nezamabadi, and S. Daly, “Perceptual signal coding for more efficient usage of bit codes,” *SMPTE Motion Imaging Journal*, vol. 122, no. 4, pp. 52–59, May 2013.

[10] *Recommendation H.264, Advanced video coding for generic audiovisual services, Series H: Audiovisual and Multimedia Systems, Infrastructure of audiovisual services – Coding of Moving Video*, ITU-T Std., 2014.

[11] *Recommendation H.265, High efficiency video coding, Series H: Audiovisual and Multimedia Systems, Infrastructure of audiovisual services – Coding of Moving Video*, ITU-T Std., 2015.

[12] J.-B. Lee and H. Kalva, *The VC-1 and H.264 Video Compression Standards for Broadband Video Services*. Springer, 2008, ch. 3, pp. 192–195.

[13] K. R. Rao, D. N. Kim, and J. J. Hwang, “Video coding standards: AVS China, H.264/MPEG-4 PART 10, HEVC, VP6, DIRAC and VC-1,” *Springer*, 2014.

[14] *VC-1 Compressed Video Bitstream Format and Decoding Process*, SMPTE Std., 2006.

[15] V. K. Goyal, “Theoretical foundations of transform coding,” *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 9–21, Sep. 2001.

[16] HEVC Test Model (HM) 14.0. Joint Collaborative Team on Video Coding (JCT-VC). https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-14.0/

[17] *Recommendation BT.2020-2, Parameter values for ultra-high definition television systems for production and international programme exchange*, ITU-R Std., Oct. 2015.

Appendix A
Proof of Lemma 1

Proof. For simplicity, define the quantizer function $\mathcal{Q}_i(x) = \mathrm{iQ}_i\left(\mathrm{Q}_i(x)\right)$. Denote the predicted frame $\mathrm{pred}(\hat{\mathbf{I}}_{t-1})$ as $\mathbf{J}_t$, and it can be decomposed into the residue vector with the smallest absolute value for each coordinate, and a vector of integer multiples of q_1, namely,

$$\mathbf{J}_t = \mathbf{J}_t \% q_1 + \mathcal{Q}_1(\mathbf{J}_t) \tag{11}$$

where $\%$ is the modulo operation. Following Fig. 1 (a), the error due to the joint effect of baseband quantization and video compression $\hat{\mathbf{I}}_t^{\text{HDR}} - \mathbf{I}_t^{\text{HDR}}$ can be written as:

$$\mathcal{Q}_1\big(\mathbf{T}^{-1}\mathcal{Q}_2\big\{\mathbf{T}\left[\mathcal{Q}_1(\mathbf{I}_t^{\text{HDR}}) - \mathbf{J}_t\right]\big\} + \mathbf{J}_t\big) - \mathbf{I}_t^{\text{HDR}}. \tag{12}$$

Substituting Eqn. (11) into (12) and moving $\mathcal{Q}_1(\mathbf{J}_t) \in (q_1\mathbb{Z})^L$ into and out of the quantizer with step size q_1, we obtain:

$$\begin{aligned}\mathcal{Q}_1\Big(\mathbf{T}^{-1}\mathcal{Q}_2\big\{\mathbf{T}\left[\mathcal{Q}_1(\mathbf{I}_t^{\text{HDR}} - \mathcal{Q}_1(\mathbf{J}_t)) - \mathbf{J}_t\%q_1\right]\big\}& \\ + \mathbf{J}_t\%q_1\Big) - \left[\mathbf{I}_t^{\text{HDR}} - \mathcal{Q}_1(\mathbf{J}_t)\right].&\end{aligned} \tag{13}$$

Here, $\mathbf{I}_t^{\text{HDR}} - \mathcal{Q}_1(\mathbf{J}_t)$ can be considered as an intra- or inter-prediction residue, and we define it as $\mathbf{r}_t^{\text{HDR}}$. In terms of quantifying error for reconstructed HDR frames, Fig. 1 (a) is therefore equivalent to Fig. 1 (b) visualized from Eqn. (13), namely,

$$\hat{\mathbf{I}}_t^{\text{HDR}} - \mathbf{I}_t^{\text{HDR}} = \hat{\mathbf{r}}_t^{\text{HDR}} - \mathbf{r}_t^{\text{HDR}}. \tag{14}$$

Assuming the quantization step of Q_1 is much smaller comparing to the range of $\mathbf{r}_t^{\text{HDR}}$, the predictive branch $\mathbf{J}_t\%q_1$ can be removed to obtain a slightly perturbed residue $\tilde{\mathbf{r}}_t^{\text{HDR}}$. Therefore, the error of non-predictively coded residues $\tilde{\mathbf{r}}_t^{\text{HDR}} - \mathbf{r}_t^{\text{HDR}} \approx \hat{\mathbf{I}}_t^{\text{HDR}} - \mathbf{I}_t^{\text{HDR}}$. □

Appendix B
Derivation for $\mathbb{E}\left[d^2\{R^{xy}[I,J]\}\right]$

Define a residue function $g(x) = \mathrm{round}(x) - x$, where $g(x) \in (-\frac{1}{2}, \frac{1}{2}]$ for any $x > 0$, and $[-\frac{1}{2}, \frac{1}{2})$ for any $x < 0$. Hence, Eqn. (5b) can be simplified to the sum of three terms:

$$\begin{aligned}\hat{\mathbf{m}} = q_1\begin{bmatrix} i \\ j\end{bmatrix} + q_2\,\mathbf{T}^{-1}g\left(\frac{q_1}{q_2}\,\mathbf{T}\begin{bmatrix} i \\ j\end{bmatrix}\right)& \\ + q_1\; g\left\{\frac{q_2}{q_1}\,\mathbf{T}^{-1}g\left(\frac{q_1}{q_2}\,\mathbf{T}\begin{bmatrix} i \\ j\end{bmatrix}\right)\right\}.&\end{aligned} \tag{15}$$

Denote the nth row and column of matrix $\mathbf{T}$ by $\mathbf{v}_n^T$ and $\mathbf{u}_n$, respectively. Define $\mathbf{p} = (i,j)$, $Y_n = g\left(\frac{q_1}{q_2}\,\mathbf{v}_n^T\mathbf{p}\right)$, and $W_n = g\left(\frac{q_2}{q_1}\,\mathbf{u}_n^T\mathbf{Y}\right)$. The squared distance is

$$d^2\{R^{xy}[I,J]\} = \|\mathbf{m} - \hat{\mathbf{m}}\|^2 \tag{16a}$$

$$= \left\|q_2\,\mathbf{T}^{-1}\begin{bmatrix} Y_1 \\ Y_2\end{bmatrix} + q_1\begin{bmatrix} W_1 \\ W_2\end{bmatrix}\right\|^2 \tag{16b}$$

$$= q_2^2\,\|\mathbf{Y}\|^2 + q_1^2\,\|\mathbf{W}\|^2 + 2q_2q_1\mathbf{Y}^T\mathbf{T}\mathbf{W}. \tag{16c}$$

Since vector $\mathbf{p} \in \mathbb{Z}_M^2$, and the term $\frac{q_1}{q_2}\mathbf{v}_n^T\mathbf{p}$ can take values on a non-degenerated subset of $\mathbb{R}$, except in very rare cases with a certain combination of $q_1, q_2, \mathbf{v}_n$ the term takes value on a subset of $\mathbb{Z}$. For the non-degenerated case, it can be proved that Y_n is approximately uniformly distributed on $(-\frac{1}{2}, \frac{1}{2})$. Therefore, $\mathbb{E}[\|\mathbf{Y}\|^2] = \frac{2}{12}$.

When $q_1 < \frac{q_2}{2}$, the range of every coordinate of $\frac{q_2}{q_1}\mathbf{Y}$ is larger than $(-1, 1)$. It can be proved that, W_n is uniformly distributed on $(-\frac{1}{2}, \frac{1}{2})$, and $\mathbf{W}$ and $\mathbf{Y}$ are uncorrelated. Therefore, $\mathbb{E}[\|\mathbf{W}\|^2] = \frac{2}{12}$, and $\mathbb{E}[\mathbf{Y}^T\mathbf{T}\mathbf{W}] = \mathrm{trace}\{\mathbf{T}\,\mathbb{E}[\mathbf{W}\mathbf{Y}^T]\}$ = 0.

In the other case, as q_1 increases, W_n becomes more depend on $\mathbf{Y}$. Statistics $\gamma_1 = \frac{12}{N}\,\mathbb{E}[\|\mathbf{W}\|^2]$ and $\gamma_{12} = \frac{12}{N}\,\mathbb{E}[\mathbf{Y}^T\mathbf{T}\mathbf{W}] = \frac{12}{N}\,\mathrm{trace}\{\mathbf{T}\,\mathbb{E}[\mathbf{W}\mathbf{Y}^T]\}$ are empirically measured, and results are shown in Fig. 3 (a).

Therefore,

$$\begin{aligned}&\mathbb{E}\big[d^2\{R^{xy}[I,J]\}\big] = \\ &\quad\begin{cases}\left(q_2^2 + q_1^2\right)/\,6, & 0 < q_1 \le \frac{q_2}{2}, \\ \left(q_2^2 + \gamma_1 q_1^2 + 2\gamma_{12}q_2q_1\right)/\,6, & \frac{q_2}{2} < q_1 \le q_2.\end{cases}\end{aligned} \tag{17}$$

And it is not difficult to generalize the above result to the N-d scenario as follows:

$$\begin{aligned}&\mathbb{E}\big[d^2\{R^{xy}[I_1, \cdots, I_N]\}\big] = \\ &\quad\begin{cases}N\left(q_2^2 + q_1^2\right)/\,12, & 0 < q_1 \le \frac{q_2}{2}, \\ N\left(q_2^2 + \gamma_1 q_1^2 + 2\gamma_{12}q_2q_1\right)/\,12, & \frac{q_2}{2} < q_1 \le q_2.\end{cases}\end{aligned} \tag{18}$$