Special Paper
방송공학회논문지 제24권 제7호, 2019년 12월 (JBE Vol. 24, No. 7, December 2019)
https://doi.org/10.5909/JBE.2019.24.7.1199
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Effective Video Compression with two Quantization Parameters

Jinwuk Seok[a)‡], Seunghyun Cho[a)], and Jin Soo Choi[a)]

Abstract

In this paper, we propose the novel methodology to improve the rate-distortion characteristics by the difference of the quantized DCT coefficients using two different quantized parameters in HEVC. Under the special condition of the quantization property in HEVC, we evaluate the binary difference of the quantized DCT coefficients derived from two different quantized parameters and compressed it with simple CABAC algorithm based on the characteristics of the binary difference. The experimental result shows that the proposed algorithm improves the compression performance of the rate-distortion property in comparison to the conventional HEVC compression method.

Keyword : HEVC, Quantization, Video compression

Ⅰ. Introduction

The issue of the quantization for a DCT(Discrete Cosine Tangent) coefficients in a video codec has been a traditional and crucial problem in video compression. In the HEVC (High Efficient Video Codec), as the newest ITU-T standard of video compression, 6 binary signal flags are employed to compressed the DCT coefficients more effectively so as to solve the mentioned issue[1]. In spite of which all techniques for the issue of the quantization of the DCT coefficients in the video compression show improved performance, it is still important to reduce the number of bits in the DCT coefficients describing the video without serious degradation of quality[2][3][4][5].

In the viewpoint of improving the coding efficiency in the context coding, we propose a novel coding technique about quantized DCT coefficients. In standard video codecs such as HEVC, for the purpose of obtaining the coding efficiency in DCT coefficients, it is necessary to code a difference between levels of quantized DCT coefficients described by several level flags. As similarly, we propose the coding methodology based on the properties of the quantization in HEVC using two different quantization parameters.

If an original video is compressed under a specific quan-

a) Electronics and Telecommunications Research Institute (ETRI)
‡ Corresponding Author : Jinwuk Seok
E-mail: jnwseok@etri.re.kr
Tel: +82-42-860-6365
ORCID: https//orcid.org/0000-0001-5318-1237

※ This work was supported by Institute for Information and communications Technology Promotion(IITP) grant funded by the Korea government (MSIP) (2017-0-00072, Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera-media).
· Manuscript received September 30, 2019; Revised November 20, 2019; Accepted November 26, 2019.

tization parameter, we derive new quantized DCT coefficients with another quantization parameter and evaluate the difference between two quantized coefficients. When we select an appropriate quantization parameter with respect to the original quantization parameter, the difference of the quantized coefficients is generated to a binary form, which is easily coded by a compressing algorithm such as a simple CABAC(Context-Based Adaptive Binary Arithmetic Coding).

As a result, it is possible to improve video quality with fewer bits in comparison to a conventional compressing technique.

This paper is organized as follows. Section 2 discusses the fundamental property of the quantization parameters in HEVC. In Section 3, we present the methodology of the implementation based on our main idea. Section 4 provides the result of computer simulation. In Section 6, we conclude this paper.

II. Fundamental Theory

1. Fundamental Property of quantization

Suppose that an input image is defined as $X \in R^{(n \times n)}$, and a prediction image evaluated by a video codec is X_t^p where the subscript denotes a time index. Under this assumption, we set a residual image X_t^e such that

$$X_t^e = X_t - X_t^p. \tag{1}$$

Therein, we define the DCT of the residual as following matrix form:

$$T(X_t^e) = X_t = T X_t^e T^T. \tag{2}$$

Herein, we set the quantization as the function of the quantization parameter q and DCT such that

$$\begin{aligned} Q(x,q) &= \left[\frac{1}{q^s}x\right] = \frac{1}{q^s}x + N_q \\ q^s \cdot \ Q(x,q) &= q^s \cdot \ \left[\frac{1}{q^s}x\right] = x + q^s \cdot \ N_q \end{aligned} \tag{3}$$

where $q^s \in \mathrm{R}$ is a scale factor for the DCT coefficients as the function of quantization parameter, $N_q \in \mathrm{R}$ is the quantization error at q such as $-1 < N_q \leq 0$, and the symbol [] is the round symbol such as $[x] \in \mathrm{Z}$ for all $x \in \mathrm{R}$ defined as follows[6][7]:

$$\forall x \in \mathrm{R}, \ [x] = \left| x + \frac{1}{2}q^s \right| \in \mathrm{Z}. \tag{4}$$

Since $N_q \leq 0$, for convenience, we introduce a positive quantization error $\epsilon_q = -N_q$ such that $0 \leq \epsilon_q < 1$. Subsequently, the quantization of the DCT coefficients is illustrated as follows:

$$\begin{aligned} Q(x,q) &= \left[\frac{1}{q^s}x\right] = \frac{1}{q^s}x - \epsilon_q \\ &= \frac{1}{q^s} T X_t^e T^T - \epsilon_q \in \mathrm{Z} \end{aligned} \tag{5}$$

Considering the inverse quantization, we can find a Quantization error occurred as the proportion to the scale factor or the quantization parameters from the following equation.

$$\begin{aligned} (Q^{-1} \circ \ Q)(x,q) &= q^s \cdot \ Q(x,q) = q^s \left[\frac{1}{q^s}x\right] \\ &= x - q^s \cdot \ \epsilon_q = T X_t^e T^T - q^s \cdot \ \epsilon_q \in \mathrm{Z} \end{aligned} \tag{6}$$

2. Quantization Error and Estimation Model

As the equation (6) illustrates that, for improvement of video quality, we should use a smaller quantization parameter. However, since the smaller quantization parameter generates more bits in compression for the DCT coefficients, an efficient coding scheme for decreasing the bits to the DCT coefficients is required.

In the sense of which an efficient coding is an encoding of a difference data between a standard signal and the other data, we set the quantized DCT coefficients by a general quantization parameter as standard data. In addition, we set the other quantized DCT coefficients as comparative data, by another quantization parameter. When we set a standard quantization parameter as q_2 and we set the other to be q_1, the quantized DCT parameters can be denoted abbreviately as $Q(X,q_2) \triangleq Q_{q_2} \in \mathrm{R}^{n \times n}$ and $Q(X,q_1) \triangleq Q_{q_1} \in \mathrm{R}^{n \times n}$.

According to the HEVC standard, the relation of the scale factor and the quantization parameter q is represented as follows [1]:

$$q^s = 0.625 \cdot 2^{(\frac{q}{6} + k \cdot (q \bmod 6))} \tag{7}$$

where mod is a modulo operation, and k is a proportional constant approximated to 0.17.

Generally, the quantization of DCT coefficients is defined such that

$$Q(X,q) = [\frac{1}{q^s} X] \approx [\frac{1}{q^s} X \times 2^{qbits}] \gg qbits \tag{8}$$

where $qbits$ is the constant value which is equal to 14, as defined in the HEVC standard , is an additional 5-bit shift considering the coding of DCT and the bit depth, which is also defined in the HEVC.

From (7) and (8), we can obtain the following approximated equation:

$$\begin{aligned} &Q(X,q) \\ &\approx \left[\frac{1}{q_m^s} \cdot 2^{qbits + \lfloor \frac{q}{6} \rfloor} \cdot \bar{X} \right] \gg (qbits + \left\lfloor \frac{q}{6} \right\rfloor + 5) \\ &= \left\lfloor \frac{1}{q_m^s} \cdot 2^{qbits + \lfloor \frac{q}{6} \rfloor} \cdot \bar{X} + f \right\rfloor \gg (qbits + \left\lfloor \frac{q}{6} \right\rfloor + 5) \end{aligned} \tag{9}$$

where f is a round factor defined in the HEVC standard.

Therefore, by (9), when there exists a difference of 6 between two quantization parameters q_1, q_2, the scale of DCT coefficients is twice difference such as

$$\left| Q_{q_1} \right| = 2 \left| Q_{q_2} \right| \ \text{for } q_1 = q_2 - 6 \tag{10}$$

However, practically, the DCT coefficients in video codec are integer values, so that there exists the quantization error represented in (3)(5). For $q_1 < q_2$, and $q_2 = q_1 + 6$ i.e. $q_1^s = \frac{1}{2} q_2^s$, suppose that $\| Q_1 \| \geq \| Q_2 \|$. Considering the quantization error, we can evaluate the error model such that

$$\begin{aligned} \Delta Q_{q_1,q_2} &= Q_1 - 2 \cdot Q_2 \\ &= \frac{1}{q_1^s} X + N_{q_1} - \frac{2}{q_2^s} X - 2N_{q_2} \\ &= \frac{1}{q_1^s} X + N_{q_1} - \frac{2}{2q_1^s} X - 2N_{q_2} \\ &= N_{q_1} - 2N_{q_2}. \end{aligned} \tag{11}$$

Since $-1 < N_q \leq 0$, we can obtain

$$-1 < N_{q_1} - 2N_{q_2} < 2 \tag{12}$$

Thereby, since $\Delta Q_{q_1,q_2} \in \mathrm{Z}$, the difference $\Delta Q_{q_1,q_2}$ is only equal to $\{0,1\}$ (in case of positive. for negative, it is equal to $\{-1,0\}$).

On the other hand, for $\| Q_1 \| \leq 2 \| Q_2 \|$, as like (11), since the difference $\Delta Q_{q_1,q_2}$ is equal to 1,

$$\begin{aligned} \Delta Q_{q_1,q_2} &= Q_1 - 2 \cdot Q_2 = \frac{1}{q_1^s} X + N_{q_1} - \frac{2}{q_2^s} X - 2N_{q_2} - 1 \\ &= \frac{1}{q_1^s} X + N_{q_1} - \frac{2}{2q_q^s} X - 2N_{q_2} - 1 = N_{q_1} - 2N_{q_2} - 1, \end{aligned} \tag{13}$$

Consequently,

$$-2 < N_{q_1} - 2N_{q_2} < 1 \tag{14}$$

,so that $\Delta Q_{q_1,q_2} \in \{0,-1\}$. The detailed proof is provided in the appendix, which is shown as theorem Ⅵ.1.

Ⅲ. Implementation of the proposed algorithm

1. Coding Structure

To provide two DCT coefficient sets under two different quantization parameter, we implement an operational function corresponding the DCT, quantization, inverse DCT and the inverse quantization. In addition, we make additional buffers for the reconstruction image, the CBF(Coded block flag) and the DCT coefficients. The concept schematic of encoder is shown in the Fig 1.

In our implementation, we set the standard quantization parameter is larger than the other quantization parameter so that the main result of compression such as the CU(Computational Unit) and PU (Prediction Unit) size is optimized to the standard quantization parameter. Consequently, the proposed algorithm does not provide the same encoding results in comparison

to the coding result only with the . Moreover, when the absolute distortion value with the standard quantization parameter is less than those of the other quantization parameter, we select the coding results of the standard quantization parameter such as the quantized DCT coefficients, CBF, and the quantization parameter itself. Finally, when there is not any compressed data such as all zero CBF, we select the reference quantization parameter which is derived from the quantization parameters in the neighbor CUs. As a result, the general rate-distortion property of the proposed algorithm follows that of the standard quantization parameter .

2. Simple CABAC coding

As illustrated in the theorem Ⅵ.1, since the difference between the quantized DCT coefficients has three state value such as -1, 0, 1, we divided the difference set to make a binary difference space. The one space includes 0 and 1 data, and the other includes common 0 and -1 data. Consequently, it is possible to represent three state difference to be a two binary difference space, and we can employ an arithmetic binary coder to code the difference data on such spaces. In contrast to the other signals in the HEVC, the probability of which the value of data represents 0 or 1 is different from the other context models, so that we should set a simple model for arithmetic coding. Considering the distribution of the difference, since the dif-

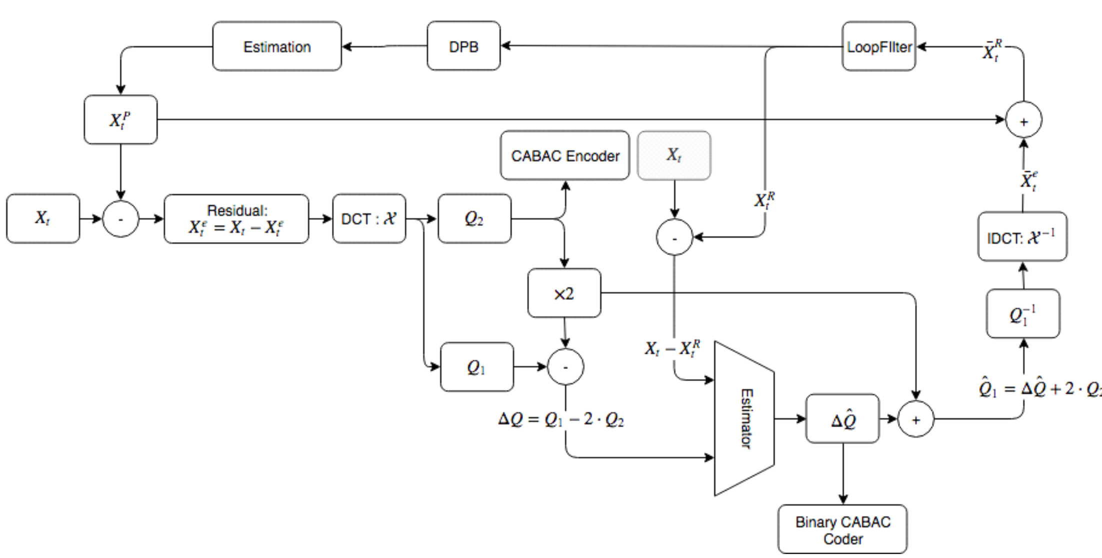

Fig. 1. Encoder Structure for the proposed algorithm

ference $\Delta Q_{q_1,q_2}$ is based on the DCT, the event of the $x_1(k)=1$ occurs more frequently as the k is small. Therefore, as the length of the code-word is relatively shorter, the probability of the event is similar to 0.5. Conversely, when those of the code-word is relatively longer, the probability approaches the 0.25. However, even though the length is relatively long, it is possible that there exists a lot of "1" in a unit code-word. Consequently, we set three probability models for the arithmetic coder. One is the simple model that the probability of "1" is initially 0.5, and it is updated by the number of "1" or "0" in a code-word as follows[7][8]:

$$p_{x_1(k)} = \frac{\Sigma_k x_1(k) + \Delta}{\Sigma_k x_0(k) + \Sigma_k x_1(k) + 2\Delta} \tag{15}$$

, where $x_y(k) \in \{0,1\}$ is a random variable to represent the kth difference data in a TU(Transform Unit) of a unit CU(Computation Unit) for $yin\{0,1\}$, and Δ is a positive integer value for the initial probability. Besides, it controls the rate of probability updating.

Another model is the complex model that the probability of "1" is defined as follows:

$$P_{x_1(k)} = \frac{E(\Sigma x_1(k)) - \Sigma_k x_1(k) + \Delta}{L - \Sigma_k x_0(k) - \Sigma_k x_1(k) + 2\Delta} \tag{16}$$

, where $E(\Sigma x_y(k))$ is the expectation sum for y in a unit code-word which is evaluated with the probability from the Theorem Ⅵ.1, and the L is the length of the code-word such as $L \triangleq (\Sigma x_0(k)) + E(\Sigma x_1(k))$. The other model is the complex model of which the probability is the same as that of the simple model.

To improve the coding efficiency and to avoid additional indication bits to the models, we set the criteria for the selection of the probability model to the arithmetic coding with analysis of the DCT/Q about the quantization parameter q_2. When the length of the nonzero signal in the DCT/Q data for q_2 is less than 5, we employ the simple model for the binary arithmetic coder. When the length of the nonzero signal in the DCT/Q data for q_2 is less than 5, we employ the simple model for the binary arithmetic coder. On the other hand, when the length is longer than 5, and the number of the larger than 2 signal for is less than 6 in a 4×4 unit TU, we set the complex model for coding, otherwise, we select the complex model with the probability of the simple model.

Ⅳ. Experimental Results

We have implemented the proposed method using HEVC reference encoder HM 15. The test sequences used in the experiments are four HD1080P (1920x1080, 4:2:0, 8bit) videos of 32 frames. For the simple probability model of the asthmatic coding, the parameter is 1. For the com-

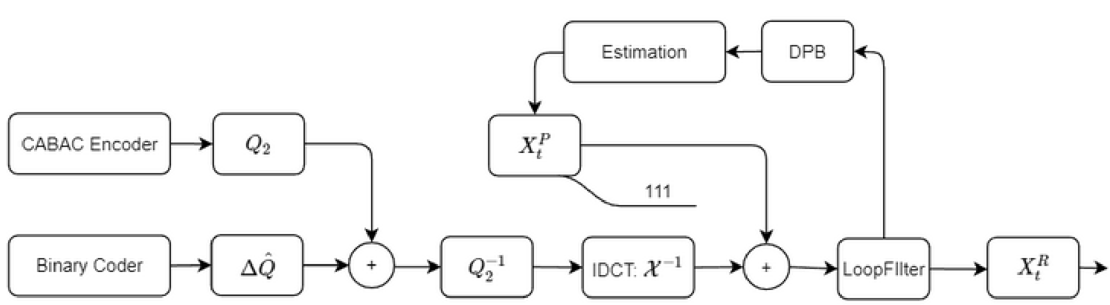

Fig. 2. Decoder Structure for the proposed algorithm

plex model with simple probability, we set it as 12. In complex model, the $E\sum x_0(k)$ is 12, and the $E\sum x_0(k)$ is 4, respectively. In addition, we set Δ to be 4 for the complex model.

Moreover, we use a rate-distortion optimization for the coding of the difference. For some CU's, by the effect of the rate-distortion optimization in DCT/Q for each quantization parameters, it is possible that the rate-distortion cost of the q_2 is less than that of the proposed algorithm. At those case, we select the result of DCT/Q for as the best result, instead of the proposed method.

The experiments are achieved under the all-Intra encoding configuration for 32 frames, and other encoding parameters, such as the size of CU or QuadTreeTU and others, are the same as the encoding parameters for the HEVC common test condition[9] with two modifications: AMP (Asymmetric Partition) is turned off and SAO(Sample Adaptive Offset) is selectively turned off. Since the proposed method employs two quantization parameters, we expect that the edge filtering between units would be more effective, if processing units are coded by different quantization parameters. However, to confirm the validity of the proposed method without such filtering effects, we skip the SAO and AMP.

As mentioned above, we apply the proposed algorithm only to the I-frames and a luminance component. The experimental results show that the improvement of BD-rate(Bjøntegaard delta bit rate) averagely about 1.2% when we use constant QP.

However, as shown in the Table 1, some experimental results with constant quantization parameters represent the large difference to bitrates between the proposed method and the anchor. Since those large differences of bitrate can distort the experimental results, we achieve another experiment with rate control appropriate to the bitrate of the proposed method. The Table 2 shows that the improvement of BD-rate averagely about 0.0217% with similar bitrates.

Table 1. Experimental Results to 4 HD1080P Test Contents with Constant QP

Sequence No.	Sequence Name	QP for I-Slice	Anchor		Proposed		
			kbps	Y psnr	kbps	Y psnr	BD-rate Y
T01	Kimono	22	77866.56	41.0402	113148.9	43.4358	-0.7%
		27	34748.64	37.9992	47482.86	39.2696	
		32	18898.08	35.6749	25269.0	36.7364	
		37	10306.56	33.2275	13476.72	34.2394	
T02	Park run	22	24120.72	43.3276	32528.82	44.0069	-2.3%
		27	13790.16	42.0175	17881.38	42.7912	
		32	8432.16	40.2343	10796.61	41.2938	
		37	5181.36	37.8772	6738.87	39.1499	
T03	Cactus	22	66913.92	41.7943	92622.6	43.4186	-1.1%
		27	36183.6	38.8595	50405.73	40.4918	
		32	19061.52	36.0117	26243.22	37.4569	
		37	9616.56	33.3189	13189.62	34.4811	
T04	Basketball Drive	22	35863.68	41.6553	49961.07	42.9024	-0.6%
		27	14378.88	39.6498	18785.31	40.2731	
		32	7529.52	38.0888	9243.36	38.6055	
		37	4250.16	36.2559	5048.34	36.7552	
Average BD-rate							-1.2%

Table 2. Experimental Results to 4 HD1080P Test Contents with Similar Bitrates

Sequence No.	Sequence Name	Anchor		Proposed		
		kbps	Y psnr	kbps	Y psnr	BD-rate Y
T01	Kimono	110123.8	43.4316	113148.9	43.4358	-0.087%
		46587.54	39.1692	47482.86	39.2696	
		24985.84	36.6049	25269.0	36.7364	
		12806.56	34.1275	13476.72	34.2394	
T02	Park run	31869.84	43.9721	32528.82	44.0069	-0.164%
		16814.57	42.6053	17881.38	42.7912	
		10298.13	41.1087	10796.61	41.2938	
		6410.82	39.0172	6738.87	39.1499	
T03	Cactus	96954.37	43.7943	92622.6	43.4186	-0.061%
		52183.6	40.6124	50405.73	40.4918	
		24837.26	37.3617	26243.22	37.4569	
		14816.38	34.5121	13189.62	34.4811	
T04	Basketball Drive	50162.25	43.0017	49961.07	42.9024	0.226%
		14378.88	39.6498	18785.31	40.2731	
		8985.67	38.4291	9243.36	38.6055	
		4871.53	36.4306	5048.34	36.7552	
Average BD-rate						-0.0217%

Ⅴ. Conclusion

We propose the novel quantization scheme for video compression using two quantization parameters in this paper. Using the property of the scale factor depending on the quantization parameter, we provide an effective coding methodology by generating a binary difference of two quantized DCT coefficients. Experimental results represent that the proposed algorithm improves the rate-distortion property in comparison to the conventional HEVC video compression. Even though we employ a simple CABAC compression arisen from the characteristics of the proposed algorithm to compress the binary difference, there exist more techniques to encode the DCT coefficients more effectively. For instance, we suppose some technique using appropriate conventional HEVC signals for encoding the DCT coefficients to decrease the context bits. Moreover, it is possible to optimize the rate-distortion in the coding of the binary difference using artificial intelligence. In addition, we will apply the proposed algorithm to Inter frames and we will extend it to obtain more efficient coding performance, through the conventional signals for DCT compression.

Ⅵ. Appendix

Herein, we provide the detailed theorem and the proof for the main analysis of the proposed method.

1. Propositions

Theorem Ⅵ.1 Let $q_2 = q_1 + 6$, i.e. $q_1^s = \frac{1}{2} q_1^s$. When Q_1 is even, $\Delta Q_{q_1 q_2} = Q_1 - 2Q_2 \in \{0,1\}$, and Q_1 is odd, $\Delta Q_{q_1, q_2} = Q_1 - 2Q_2 \in \{0,-1\}$.

Proof.

Let $X = q_1 \cdot k + m$ where $0 \leq m < q_1$, then

$$\left[\frac{1}{q_1}X\right]=\left[k+\frac{1}{q_1}m\right]=\left\lfloor k+\frac{1}{q_1}m+\frac{1}{2}\frac{1}{q_1}q_1\right\rfloor. \qquad (17)$$

Thereby,

$$\left[\frac{1}{q_1}X\right]=\begin{cases} k & 0\le m\frac{1}{2}q_1 \\ k+1 & \frac{1}{2}q_1\le m<q_1. \end{cases} \qquad (18)$$

For q_2, by the same way, we can obtain

$$\left[\frac{1}{q_2}X\right]=\left[\frac{1}{2q_1}X\right]=\left[\frac{1}{2}k+\frac{1}{2q_1}m\right]=\left\lfloor \frac{1}{2}k+\frac{1}{2}\frac{1}{2q_1}2q_1\right\rfloor \qquad (19)$$

Let $k=2\cdot\bar{k}+n$. In (19), since $m<q_1$, $\frac{1}{2}<\frac{m}{2q_1}+\frac{1}{2}<1$, we can obtain

$$\left[\frac{1}{q_2}X\right]=\begin{cases} \left\lfloor \frac{1}{2}k+\frac{1}{2q_1}m+\frac{1}{2}\right\rfloor=\bar{k}. & n=0 \\ \left\lfloor \frac{1}{2}k+\frac{1}{2}+\frac{1}{2q_1}m+\frac{1}{2}\right\rfloor=\bar{k}+1 & n=1 \end{cases} \qquad (20)$$

The equation (20) means that, if the k is odd number, $\left[\frac{1}{q_2}X\right]$ is $\bar{k}+1$ in spite of which value the remainder m has.

- When k is odd value and $m<\frac{1}{2}q_1$,

$$\left[\frac{1}{q_1}X\right]=2\bar{k}+1,\ \left[\frac{1}{q_2}X\right]=\bar{k}+1. \qquad (21)$$

Therefore,

$$Q_1-2Q_2=2\bar{k}+1-2(\bar{k}+1)=-1 \qquad (22)$$

When k is odd value and $\frac{1}{2}q_1\le m<q_1$,

$$\left[\frac{1}{q_1}X\right]=(2\bar{k}+1)+1,\ \left[\frac{1}{q_2}X\right]=\bar{k}+1. \qquad (23)$$

Therefore,

$$Q_1-2Q_2=2Q_2=2\bar{k}+2-2(\bar{k}+1)=0 \qquad (24)$$

When k is even value and $m<\frac{1}{2}q_1$,

$$\left[\frac{1}{q_1}X\right]=2\bar{k},\ \left[\frac{1}{q_2}X\right]=\bar{k}, \qquad (25)$$

Therefore,

$$Q_1-2Q_2=2\bar{k}-2\bar{k}=0 \qquad (26)$$

When k is even value and $\frac{1}{2}q_1\le m<q_1$

$$\left[\frac{1}{q_1}X\right]=2\bar{k}+1,\ \left[\frac{1}{q_2}X\right]=\bar{k}, \qquad (27)$$

Therefore,

$$Q_1-2Q_2=2\bar{k}+1-2\bar{k}=1. \qquad (28)$$

■

Theorem Ⅵ.2 The fundamental equation of quantization parameter is defined as the equation (7) and (8), the quantization of DCT is approximated as follows:

$$\begin{aligned} &Q(X,q) \\ &\approx\left[\frac{1}{q_m^s}\cdot 2^{qbits+\left\lfloor\frac{q}{6}\right\rfloor}\cdot\bar{X}\right]\gg\left(qbits+\left\lfloor\frac{q}{6}\right\rfloor+5\right) \\ &=\left\lfloor\frac{1}{q_m^s}\cdot 2^{qbits+\left\lfloor\frac{q}{6}\right\rfloor}\cdot\bar{X}+f\right\rfloor\gg\left(qbits+\left\lfloor\frac{q}{6}\right\rfloor+5\right) \end{aligned} \qquad (29)$$

Proof.

$$
\begin{aligned}
&Q(X,q) \\
&\approx [\frac{1}{0.625 \cdot 2^{\lfloor \frac{q}{6} \rfloor + k \cdot m}} X \times 2^{qbits}] \gg qbits \\
&= [\frac{8}{5} \cdot 2^{-(\lfloor \frac{q}{6} \rfloor + k \cdot m)} X \times 2^{qbits}] \gg qbits \\
&= [\frac{8}{5} \cdot 2^{qbits - k \cdot m} X \times 2^{-\lfloor \frac{q}{6} \rfloor}] \gg qbits \\
&\approx [\frac{8}{5} \cdot 2^{qbits - (\log_2 \frac{q_m^s}{0.625} - \lfloor \frac{q}{6} \rfloor)} X] \gg (qbits + \lfloor \frac{q}{6} \rfloor) \\
&= [(\frac{8}{5} \cdot 2^{qbits - \log_2 \frac{q_m^s}{0.625}}) 2^{\lfloor \frac{q}{6} \rfloor} X] \gg (qbits + \lfloor \frac{q}{6} \rfloor) \\
&= [(\frac{8}{5} \cdot 2^{-\log_2 \frac{q_m^s}{0.625}} \cdot s^{qbits}) \cdot 2^{\lfloor \frac{q}{6} \rfloor} X] \gg (qbits + \lfloor \frac{q}{6} \rfloor) \\
&= [(\frac{8}{5} \cdot \frac{0.625}{q_m^s} \cdot 2^{qbits}) \cdot 2^{\lfloor \frac{q}{6} \rfloor} X \gg (qbits + \lfloor \frac{q}{6} \rfloor) \\
&= [(\frac{1}{q_m^s} \cdot 2^{qbits}) \cdot 2^{\lfloor \frac{q}{6} \rfloor} X] \gg (qbits + \lfloor \frac{q}{6} \rfloor)
\end{aligned}
\tag{30}
$$

■

2. Mathematical Formulation of Inverse Quantization

The inverse quantization is basically formulated by multiplication of the quantization step $q^s \in \mathbb{R}$ to the quantized DCT coefficients as follows.

$$
Q^{-1}(q, X^Q) = X^Q \cdot q^s \tag{31}
$$

However, since the quantization step is a real value, we have to change it as an integer value in encoding and decoding processes. To make an integer quantization step, we use it an integer scale variable and rewrite the equation such that:

$$
\begin{aligned}
Q^{-1}(q, X^Q) &= X^Q \cdot 0.625 \cdot 2^{\lfloor \frac{q}{6} \rfloor + k \cdot (q \bmod 6)} \cdot 2^6 \\
&= X^Q \cdot \frac{5}{8} \cdot 2^{\lfloor \frac{q}{6} \rfloor + k \cdot (q \bmod 6)} \cdot 2^3 \cdot 2^3 \\
&= X^Q \cdot 40 \cdot 2^{\lfloor \frac{q}{6} \rfloor + k \cdot (q \bmod 6)}
\end{aligned}
\tag{32}
$$

In the HEVC standard, the scale parameter, which is equal to 2^4, to the quantization, is added, and the value of inverse quantization is divided by 2^{shift} for the bit-depth of an input image and coding scale appropriate to the range of the DCT transformation. Moreover, since the division of 2^{shift}, it is necessary to add the term which integrase the inverse quantized value by round calculation. As a result, the final formulation of inverse quantization can be written such that:

$$
\begin{aligned}
&Q^{-1}(q, X^Q) \\
&= [X^Q \cdot 40 \cdot 2^{\lfloor \frac{q}{6} \rfloor + k \cdot m + 4}] \gg shift \\
&= \lfloor X^Q \cdot 40 \cdot 2^{\lfloor \frac{q}{6} \rfloor + k \cdot m + 4} + 2^{shift - 1} \rfloor \gg shift
\end{aligned}
\tag{33}
$$

References

[1] ISO/IEC, *High Efficiency Video Coding, document ITU-T Rec. H.265 and ISO/IEC 23008-2 (HEVC)*, ITU-T and ISO/IEC, Sep. 2019

[2] B. Bross, W.-J. Han, J.R. Ohm, G. J. Sullivan, Y.K. Wang, and T. Wiegand, "High Efficiency Video Coding (HEVC) text specification draft 10 (For FDIS & Final Call)", JCT-VC, Doc. JCTVC-L1003, Geneva, Switzerland, 2013.

[3] D. Marpe, H. Schwarz, T. Wiegand, "Context-based Adaptive Binary Arithmetic Coding in the H.264/AVC Video Compression Standard", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13, No. 7, pp.620–636, 2003.

[4] R. Sjöberg, Y. Chen, A. Fujibayashi, M.M. Hannuksela, J. Samuelsson, T.T. Keng, and Y. Wang, W. Stephan, "Overview of HEVC High-Level Syntax and Reference Picture Management", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 22, No. 12, pp.1858–1870, 2012.

[5] H. Schwarz, T. Nguyen, D. Marpe, T. Wiegand, *Non-CE7: Alternative entropy coding for dependent quantization*, JVET-K0072, Proceedings of 11th JVET meeting, Ljubljana, SI, 2018.

[6] R.M. Gray, D.L. Neuhoff, "Quantization", IEEE Transactions on Information Theory archive, Vol. 44, No. 6, pp.2325-2383, 1998

[7] R.G. Gallerger, *Information Theory and Reliable communication* , NewYork, Wiley, 1968

[8] D. L. Duttweiler and C. Chamzas, "Probability estimation in arithmetic and adaptive-Huffman entropy coders,"IEEE Transactions on Image Processing, Vol. 4, No. 3, pp.237-246, 1995.

[9] D. Taubman, M. Marcellin, ' JPEG2000 *Image Compression Fundamentals, Standards and Practice*', Springer US, NewYork, 2002

[10] F. Bossen, "Common test conditions and software reference configurations", The 8th JCT-VC meeting, JCT-VC H1100, San Jose, CA, Jan. 2012.

저 자 소 개

Jinwuk Seok

- 1993. 2. : Dept. of Electrical and Control Engineering, Hong-Ik University (B.S.)
- 1995. 2. : Dept. of Electrical and Control Engineering, Hong-Ik University (M.S.)
- 1998. 8. : Dept. of Electrical Engineering, Hong-Ik University (Ph.D.)
- 2000 ~ Current : Principal Researcher, ETRI
- 2009 ~ Current : Adjunct Professor, UST
- ORCID : https//orcid.org/0000-0001-5318-1237
- Research interests : Video Processing/Coding. Machine Learning, Nonlinear Control

Seunghyun Cho

- 2003. 8. : Dept. of Electronic and Electrical Engineering, Kyungpook National University (B.S.)
- 2006. 2. : Dept. of Electronic and Electrical Engineering, KAIST (M.S.)
- 2015. 8. : Dept. of Electronic and Electrical Engineering, KAIST (Ph.D.)
- 2006. 6. ~ Current : Principal Researcher, ETRI
- ORCID : https://orcid.org/0000-0003-1985-4420
- Research interests : Video Processing/Coding, Computer Vision, Machine Learning, Digital Circuit Design

Jin Soo Choi

- 1990. 2. : Dept. of Electronic Engineering, Kyungpook National University (B.S.)
- 1992. 2. : Dept. of Electronic Engineering, Kyungpook National University (M.S.)
- 1996. 2. : Dept. of Electronic Engineering, Kyungpook National University (Ph..D.)
- 1996. 5. ~ Current : Principal Researcher, ETRI
- ORCID : https://orcid.org/0000-0003-4297-5327
- Research interests : Video Processing/Coding, UHDTV Broadcasting, 3DTV Broadcasting