

A Fast Zero-Quantized Percentage Model for Video Coding with RDO Quantization

Haoyun Yang, Haibing Yin^(⊠), and Xiaofeng Huang

School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China 15382363770@163.com

Abstract. In video coding, the percentage of zero-quantized coefficients, denoted by ρ , is directly determined by the quantization algorithm adopted. ρ domain rate distortion (RD) modeling is widely employed to optimize the implementation algorithm for customizable modules such as rate control and mode decision etc. How to calculate or estimate ρ according to quantization algorithm is the first step task for ρ -domain RD modeling. There are two typical quantization algorithm, soft-decision quantization such as dead-zone, and softdecision quantization such as rate distortion optimized quantization (RDOQ). RDOQ is more frequently employed in the latest video encoders compared with deadzone quantization due to its inspiring coding performance. In HDO based video encoder, ρ can be easily obtained by simply rounding. However, it is computation-intensive to calculate ρ in video encoder with RDOQ, in which complicated trellis search is employed. This paper focus on developing estimation model for quickly estimating ρ for RDOQ based video coding. The contribution of this article is as follows: First, this paper develops the ρ model adaptively according to an adaptive deadzone offset model, which is modeled by imitating the behavior of RDOQ. Second, an accurate ρ model is adaptively built offline as function of weighted SATD (sum of absolute transformed distortion) denoted by WSATD, quantization step size a, and average WSATD/a estimated from ensemble. The weight in WSATD is adaptively determined according to the adaptive offset to simulate the behavior pattern of RDOQ as much as possible. Experimental results verify that the proposed model can quickly and accurately predict the ρ results of RDOQ with moderate implementation complexity. The proposed ρ model can be employed to estimate the percentage of zero quantized coefficients which can be used for fast all-zero detection and ρ domain rate distortion modeling in RDOQ based HEVC video encoder.

Keywords: Video coding · RDOQ · Rate-distortion model Percentage of zero-quantized coefficients

1 Introduction

In video coding, the rate-distortion (R-D) function model is widely employed for rate control and mode decision. Several R-D models for DCT-based video coding have been proposed in the literature [1]. Accurately calculating R and D will cost video

encoder high computation complexity especially in the latest HEVC coders with computation-intensive RDOQ and CABAC entropy codec. Developing fast estimation model for R and D had attracted intensive research interests in the past twenty years. Quantization directly determine the coding distortion and rate consumption. Thus, some works explore rate distortion models as function of quantization parameter. Some works explore the relationship between RD models with the percentage of zeroquantized coefficients denoted by ρ . ρ -domain modeling or q-domain modeling are two typical rate distortion modelling methods. Relatively, ρ -domain modeling can more accurately describe the microcosmic characteristics of CABAC in terms of accurate R estimation. Several works had also verified that ρ has a critical effect on the coding bit rate R, especially at low bit rates [2].

On the other hand, there are great amounts of blocks are quantized to all-zero in the latest HEVC standard, especially at low bit applications. In the latest standard, rate distortion optimization is applied in quantization. In RDO based quantization, quantization consumes high computation complexity due to that the quantizer need to evaluate the distortion and rate for all possible candidate results and select an optimal result in the sense of rate distortion optimization. If we can do all zero block (AZB) detection before using RDO, we can reduce the HEVC coding burden. In the past two decades, there are several AZB detection algorithms reported in the literature [3]. From the viewpoint of target quantization algorithm, these AZB decision algorithm were usually designed for video encodes with HDQ, for example dead-zone quantization, in which RD optimized RDOQ is not supported [3]. Thus, we can establish a ρ model for RDOQ to indirectly implement all-zero block decisions before using RDO.

 ρ model is used to estimate the percentage of zero-quantized coefficients, and thus it is highly related with the quantization algorithm adopted in video encoder. There are two typical quantization algorithms in prevailing video encoders, hard-decision quantization (HDQ) such as deadzone and soft-decision quantization (SDQ) such as rate distortion optimized quantization (RDOQ). In HDQ based video encoder, ρ can be easily obtained by simply rounding. However, RDOQ can achieve superior coding rate distortion performance compared with HDQ. It is computation-intensive to calculate ρ in video encoder with RDOQ, in which complicated trellis search is employed. Thus, it is meaningful to develop ρ model for RDOQ based video coding.

Hence, this paper proposes ρ model for RDOQ based video coding. Firstly, we formulate ρ as functions of quantization step size, weighted *SATD* (sum of absolute transformed difference) and the mean of *WSATD*. By accurately measuring the adaptive deadzone offset estimated from the DCT coefficient distribution parameter Λ , we define an adaptive weight model for weighted *SATD* to obtain the adaptive offset δ and apply it to the weight model. Then, the three-dimensional ρ model is constructed using statistical curve fitting method from ensemble. The ρ model is developed individually in the cases of different types of TU blocks. The proposed model can quickly and accurately predict the ρ results of RDOQ.

This paper is organized as follows. Problem formulation and motivation analysis are given in Sect. 2. The proposed ρ model is given in Sect. 3. Section 4 gives the experimental results. Section 5 concludes the whole paper.

2 Problem Formulation and Motivation

2.1 RDOQ and ρ Model

 ρ model is used to estimate the percentage of zero-quantized coefficients, and thus it is highly related with the quantization algorithm adopted in video encoder. There are two typical quantization algorithm, soft-decision quantization such as dead-zone, and softdecision quantization such as rate distortion optimized quantization (RDOQ). In HDQ based video encoder, the quantization result z_i is adjusted using a rounding deadzone offset *f* described as follows:

$$z_i = floor\left(\frac{|c_i|}{q} + f\right) \tag{1}$$

where floor(.) is a direct integer operation, c_i is the DCT coefficient, and q is the quantization step size. In deadzone HDQ, ρ can be easily obtained by simply rounding.

In RDOQ, several candidate quantization results are determined according to the results of HDQ. Then, rate distortion optimization is employed to further refine the optimal quantization result from three candidate results. In RDOQ, inter-coefficient correlation is taken into consideration by joint rate distortion optimization with context adaptive binary arithmetic coding (CABAC). Suppose there are *N* coefficient in the current transform block, and there are *m* candidate quantization results preselected for further refinement. For a specific coefficient c_i , its candidate results are l_{i1} , l_{i2} ,..., l_{im} which are centered about the result of HDD obtained with fixed rounding offset f = 0.5. RDOQ checks all candidates to select an optimal result l_i using RDO described [4] as follows.

$$l_{i} = \underset{k=1 \sim m}{\operatorname{arg\,min}} \left\{ D(c_{i}, l_{ik}) + \lambda \cdot R(l_{ik}) + \sum_{j=i+1}^{N} D(c_{j}, \overline{l_{j}}) + \lambda \cdot R(\overline{l_{j}}) \right\}$$
(2)

where $D(c_i, l_{ik})$ and $R(l_{ik})$ are the coding distortion and rate when c_i is quantized to l_{ik} , and λ is the Lagrangian multiplier, and \overline{l}_j is the initial center of all candidates, i.e. the HDQ quantization result of the *j*-th coefficient in the current block. As shown in the above equation, inter-coefficient influence is considered in RDOQ. The backward intercoefficient rate propagation is taken into consideration to optimally determine a specific coefficient's quantization result.

In general, trellis search is employed to solve this dynamic programming problem. In HEVC video codec reference model (HM), simplified trellis search is implemented to alleviate the heavy computation of the full trellis search. Nevertheless, the computation is still relatively high. As a result, it is computation-intensive to calculate ρ in video encoder with RDOQ, in which complicated trellis search is still desired instead of simple rounding. Since RDOQ can achieve superior coding rate distortion performance compared with HDQ, thus, it is meaningful to develop ρ model for RDOQ based video coding.

2.2 Analysis on RDOQ Based ρ Model

In order to develop ρ model, we need to build a function between ρ and some characteristic parameters which characterize the block and are relatively easy to be obtained for reasonable model computation complexity.

First, the sum of absolute transformed distortion denoted by *SATD* is usually employed to describe the current block's characteristics. In addition, it is also easy to obtain due to that it is available after mode decision. *SATD* was widely employed to develop the RD models in traditional works. In this work, we also tend to explore the relationship between ρ and *SATD*. In video coding, *SATD* is defined as follows.

$$SATD = \sum_{i} |c_i| \tag{3}$$

where c_i is the DCT coefficient. We can conclude that *SATD* is directly proportional to the amplitudes of all DCT coefficients as shown in Eq. (3).

We cannot directly apply *SATD* into ρ model building. In terms of ρ modeling, there is nonlinear relationship between *SATD* and ρ . For example, if one coefficient is larger than the optimal zero-quantized threshold, it is quantized to non-zero coefficient. However, increasing the coefficient intensity do not change a certain coefficient's contribution to ρ of the current block. In other words, if $|c_i|$ is larger than the optimal zero-quantized threshold, no matter how large $|c_i|$ is, the contribution of the current coefficient is stable.

What if we change the *SATD* definition accounting for the above nonlinearity in terms of ρ contribution? This paper tends to adopt weighted *SATD* to alleviate the nonlinearity extent between *SATD* and ρ . We propose a weight model according to the intensity of c_i to approximate a linear relationship between weight *SATD* and ρ .

Second, there is an intrinsic relationship between ρ and the quantization step size q. In general, the larger q is, the more coefficients are quantized to zero. There is proportional relationship between ρ and q. By jointly taking *WSATD* and q into consideration, we employ *WSATD*/q denoted by χ to build a relationship among ρ , *WSATD* and q. Suppose we apply a weight factor w_i for the coefficient c_i , we can then define the composite parameter χ as follows.

$$\chi = \sum_{i} \frac{w_i \times c_i}{q} \tag{4}$$

Where c_i is the DCT coefficient, w_i is the weight factor, and q is the quantization step size. Here, how to adaptively determine the weight factor w_i is one important problem.

Third, in our simulation, we find that the scatter results of (ρ, χ) samples are not convergent enough to use a close-form function to formulate it. RDOQ uses complicated trellis search to deal with inter-coefficient influence. It is not enough to accurately describe ρ only according to χ . Therefore, we need to introduce another parameter to develop more accurate three-dimensional ρ -model to imitate the behavior pattern of the optimal RDOQ.

By statistical analysis on the ρ - χ samples, we found that the average of χ estimated from a sliding window in the case of the same ρ , denoted by ω , also have a regular and obvious functional relationship with ρ . χ can measure the ensemble characteristics from the viewpoint of large sample analysis. Consequently, this paper will take the ω (average χ) as the third parameter for developing the ρ model. Figure 1 below shows the framework of the proposed ρ model.



Fig. 1. Framework of the proposed ρ model

3 The Proposed ρ Model

3.1 WSATD with Deadzone Offset Adaptive Weight

Judging whether one DCT coefficient is quantized to zero or not accurately is the key to building the weight model for *WSATD* as analyzed in Sect. 2. The quantization determines the weight directly in terms of *SATD* weighting. In HDQ, whether or not c_i is quantized to zero can be determined by simple rounding according to deadzone offset f. However, in the RDOQ, it is computation-intensive to determine whether c_i is quantized to zero or not accurately.

In our previous work, we had made in-depth research to model an adaptive deadzone offset δ to improve the deadzone HDQ. By imitating the behavior pattern of RDOQ using statistic analysis, the offset δ mode is modeled as function of quantization parameter, quantization remainder, and the DCT coefficient distribution parameter Λ . This model is built offline, a three-dimension table is given offline. With this model, δ can be simply estimated by simple table lookup [5]. Based on Laplacian model, the distribution of DCT coefficients Λ can be estimated as follows.

$$\Lambda = \frac{1}{n} \sum_{i=1}^{n} |c_i| \tag{5}$$

where c_i represents the DCT coefficient and n represents the number of statistical coefficients.

In general, the larger *SATD* is, the higher probability that coefficients are quantized to nonzero, and the smaller ρ is. Aiming at develop *WSATD* which is inversely proportional to ρ , this work proposes a weight model to adaptively adjust the contribution to *WSATD* according to coefficient the intensity $|c_i|$. If one coefficient c_i is quantized by RDOQ to zero, its contribution to *WSATD* is also close to zero. If one coefficient c_i is quantized to non-zero by RDOQ, its contribution to *WSATD* is identical regardless of its intensity $|c_i|$. Intuitively, the larger $|c_i|$ is, the smaller the corresponding weight w_i is.

That is, $|c_i|$ and w_i approximately comply with negative exponential function. This paper adopts the adaptive weight model shown as follows.

$$w_i = e^{-(|c_i| - b)/a} \tag{6}$$

where c_i represents the DCT coefficient, and w_i is the resulting weight. We can control the slope and centroid of the functional curve according to the $|c_i|$ by employing two control parameters *a* and *b*.

TU type	m	Total blocks	ALL-zero blocks	Percentage	Non-zero blocks	Percentage
4×4	2	159173	132871	83.48%	26302	16.52%
8×8	3	108598	100647	92.68%	7951	0.72%
16×16	6	140682	135398	96.24%	5284	3.76%
32×32	6	26594	26445	99.44%	149	0.056%

 Table 1. m numerical statistics

In order to quickly determine w_i , we can pre-judge whether c_i is quantized to zero in the case of RDOQ according to c_i . On one hand, according to the principle of HDQ, coefficients whose $|c_i|$ are within the range [0, (1 - f)q) are quantized by HDQ to zero, and these coefficients are also quantized to zero ones by RDOQ. On the other hand, for the coefficients whose $|c_i|$ are within the range [(1 - f)q, q), the coefficients are quantized to nonzero by HDQ, they may be quantized to zero or nonzero coefficients by RDOQ. In general, if one TU only contain m possible nonzero coefficients with intensity [(1 - f)q, q), the current block may be quantized to all-zero block by RDOQ in the sense of rate distortion optimization due to that quantizing all coefficients to zero will save some coding bits. We have observed some sample blocks that are quantized to all-zero by RDOQ although some coefficients are quantized to nonzero by HDQ. The typical m is given in Table 1 in the cases of different TU blocks with different block size.

As analyzed above, if c_i is quantized to non-zero by RDOQ, different $|c_i|$ intensities contribute to final ρ identically. Using this property, we propose heuristic way to determine the control parameters a and b to obtain accurate WSATD for ρ modeling. We need to deterministic control points $(|c_i|, w_i)$ for parameter selection. On one hand, suppose a coefficient with intensity $(1 - \delta)q$ have the normalized contribution 1 for WSATD estimation and ρ . As a result, $((1 - \delta)q, 1)$ is used as one control point. On the other hand, suppose that the maximal of $|c_i|$ is c_{max} , a corresponding minimal weight w_{min} is supposed to be determined. The coefficient with maximal $|c_i|$ is supposed to contribute the same degree with the coefficient with intensity $(1 - \delta)q$. As a result, another control point (c_{max}, w_{min}) is derived according to $c_{max}*w_{min} = (1 - \delta)q$. As a result, we can derive a and b according to Eq. (6) using the above two control points. a and b vary dynamically according to the TU block, resulting in accurate weight model w_i .

3.2 ρ Model

As analyzed in Sect. 2, ρ has an intrinsical functional relationship with χ and ω . This work determines a three-dimensional ρ model (denoted by ρ - ω - χ) via surface fitting. How to accurate estimate ω online is the first task here. In order to achieve fast ρ estimation, we need to build a model to estimate the averaged χ , i.e. ω .

Suppose that the percentage of zero-quantized coefficients in the case of HDQ is denoted as ρ' . Using RDO to determine the optimal quantization result, only very few coefficients with nonzero-quantized HDQ results are finally quantized to zero coefficients. As a result, ρ' is basically identical with ρ in general. Supposed that the averaged χ in the case of identical ρ' is denoted as ω' . Similarly, ω' is usually identical with ω . As a result, we can estimate ω according to ω' instantly due to that ω' can be easily obtained online.



Fig. 2. Fitting curves and scatter plots between ω - ω'

In terms of ω - ω' function modeling, we need to clean the samples collected. The first case is that ω in the case of certain ρ may not exist. In this case, we need to remove the corresponding ω' sample in the case of the corresponding ρ' (whether or not the mean exists). The second case is that ω' in the case of certain ρ' may not exist. In this case, we need to determine ω in the case of corresponding ρ with ω' (if ω exists). There is high correlation between ρ' and ω' , as a result we can develop modeling ω' - ρ' . Two function models ω' - ρ' and ω - ω' can be obtained by curve fitting respectively:

$$\omega' = \mathbf{g}(\rho') \tag{7}$$

where ρ' is the percentage of zero-quantized coefficients by HDQ based prequantization, ω' is the mean of χ in the case of identical ρ' . There is a monotonically decreasing function of the second order between ω and ω' , and this function can be formulated as follows.

$$\omega = f(\omega') \tag{8}$$

Where ω' is the mean of χ in the case of identical ρ' , ω is the mean of χ in the case of identical ρ . Figure 2 shows the fit curve and scatter plot results between ω and ω' .

According to Fig. 2, we can derive the following conclusions. When the TU block type is 4×4 and 8×8 , we can accurately predict ω according to Eqs. (7) and (8). For other types of TU blocks, Eqs. (7) and (8) cannot accurately predict ω sometimes. In order to ignore these outlier samples, we employ the estimated ω' to replace ω for ρ - ω modeling. Finally, the function model ρ - ω is obtained by curve fitting. Figure 3 below shows the fit curve and sample scatter results.



Fig. 3. Fitting curves and sample scatter results

Through Figs. 2 and 3, we can derive the parameter ω of the ρ model, which can be obtained by $\omega' - \rho'$ and $\omega - \omega'$ models online. $\rho - \omega$ has a high functional correlation. With the estimated ω , we then finally establish a three-dimensional ρ model by surface fitting, which is implemented online. As shown in the following formula (9).

$$\rho = f(\omega, \chi) \tag{9}$$

The three-dimensional ρ model is a polynomial with respect to the one order of χ and the third order of ω .



Fig. 4. Fitting surfaces and sample scatter results

Figure 4 gives the corresponding fitted surfaces and sample scatter results. Figure 4 are intensively shown from two angles of view for better understanding.

From the results in Fig. 4, we can draw the following conclusions. On one hand, the proposed ρ model can accurately predict the ρ results of RDOQ. Compared with the percentage of zero-quantized coefficients by HDQ, the prediction results obtained by the proposed model is very close to the actual ρ samples. In order to evaluate the accuracy of the proposed ρ model, we also report the estimated error ratio. The resulting estimation error ratios are given in Fig. 5, in which the histogram results of the prediction error are given.



Fig. 5. Error frequency histogram

4 Experimental Result

The proposed ρ model can quickly and accurately predict the ρ results of RDOQ. In order to evaluate the model accuracy of ρ model, we take the ρ results of RDOQ as the comparison anchor and investigate the estimation error of the proposed ρ model relative to the results of RDOQ. Here, we compare the simple ρ model estimated from simple deadzone HDQ with the proposed ρ model. The estimation error results of two ρ models are given in Table 2. There simulation results in the cases of different QP and TU block size are given.

According to the results in Table 2, we can draw the following conclusions. In the cases of different TU blocks, the proposed ρ model established off-line can quickly and accurately predict the ρ results of RDOQ, and the error results are usually smaller than 0.01. Comparatively, in the case of HDQ based ρ model, a great amount of samples suffer from estimation error larger than 0.01, and some samples have estimation error ratio close to 0.2. In terms of computational complexity, only an additional curve and surface fitting are desired. The additional complexity of proposed model is moderate.

QP	Basketba	llPass		BasketballDrill		
	TU type	Error (HDQ)	Error (model)	TU type	Error (HDQ)	Error (model)
32	4×4	0.35	0.05	4×4	0.25	0.04
	8×8	0.2	0.01	8×8	0.15	0.01
	16×16	0.13	0.005	16×16	0.06	0.005
	32×32	0.1	0.001	32×32	0.06	0.001
37	4×4	0.25	0.05	4×4	0.25	0.05
	8×8	0.2	0.01	8×8	0.15	0.01
	16×16	0.1	0.005	16×16	0.06	0.005
	32×32	0.1	0.001	32×32	0.05	0.0009
QP	BQMall			Johnny		
	TU type	Error (HDQ)	Error (model)	TU type	Error (HDQ)	Error (model)
32	4×4	0.25	0.03	4×4	0.25	0.04
	8×8	0.15	0.006	8×8	0.15	0.009
	16×16	0.1	0.001	16×16	0.05	0.003
	32×32	0.08	0.0009	32×32	0.05	0.0009
37	4×4	0.25	0.02	4×4	0.2	0.04
	8×8	0.15	0.006	8×8	0.1	0.007
	16 imes 16	0.06	0.0001	16×16	0.03	0.003
	32×32	0.05	0.0005	32×32	0.04	0.0004
QP	FourPeople			KristenAndSara		
	TU type	Error (HDQ)	Error (model)	TU type	Error (HDQ)	Error (model)
32	4×4	0.2	0.04	4×4	0.2	0.03
	8×8	0.15	0.01	8×8	0.1	0.01
	16×16	0.06	0.005	16×16	0.04	0.003
	32×32	0.03	0.001	32×32	0.03	0.0009
37	4×4	0.2	0.03	4×4	0.2	0.03
	8×8	0.1	0.01	8×8	0.1	0.01
	16×16	0.025	0.001	16×16	0.04	0.003
	32×32	0.022	0.001	32×32	0.03	0.001

Table 2. ρ estimation error results of the two models compared with RDOQ

5 Conclusion

In video encoder, the percentage of zero-quantized coefficients ρ is useful for rate distortion model building and all-zero block detection. Developing fast ρ estimation model plays important role in rate distortion optimization for video coding. This paper proposes a fast ρ model for RDOQ based video coding. An accurate ρ model is adaptively built offline as function of weighted *SATD*, quantization step size, and average *WSATD/q* estimated from ensemble. Experimental results verify that the proposed model can quickly and accurately predict the ρ results of RDOQ with moderate

implementation complexity. The proposed ρ model can be employed to optimized allzero block detection and rate distortion model building.

References

- He, Z., Kim, Y.K., Mitra, S.K.: Low-delay rate control for DCT video coding via p-domain source modeling. IEEE Trans. Circuits Syst. Video Technol. 11(8), 928–940 (2001)
- Milani, S., Celetto, L., Mian, G.A.: An accurate low-complexity rate control algorithm based on (ρ, Eq)-domain. IEEE Trans. Circuits Syst. Video Technol. 18(2), 257–262 (2008)
- Lee, K., et al.: A novel algorithm for zero block detection in high efficiency video coding. IEEE J. Sel. Topics Signal Process. 7(6), 1124–1134 (2013)
- 4. Yang, K.H.: Methods and systems for rate-distortion optimized quantization of transform blocks in block transform video coding. US, US7957600 (2011)
- Wang, H., Yin, H., Shen, Y.: A Novel Hard-Decision Quantization Algorithm Based on Adaptive Deadzone Offset Model. In: Chen, E., Gong, Y., Tie, Y. (eds.) PCM 2016. LNCS, vol. 9917, pp. 335–345. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48896-7_ 33