COMPLEXITY REDUCTION ALGORITHM FOR OPTIMUM QUANTIZER DESIGN BASED ON AMPLITUDE SPARSENESS

Yukihiro BANDOH, Seishi TAKAMURA, and Atsushi SHIMIZU

NTT Media Intelligence Laboratories, NTT Corporation
1-1 Hikari-no-oka, Yokosuka, Kanagawa, 239-0847 JAPAN
E-mail : bandou.yukihiro@lab.ntt.co.jp

ABSTRACT

The design of an optimum quantizer can be formulated as an optimization problem that finds the quantization indices that minimize the quantization error. One solution of the optimization problem is DP quantization, an approach based on dynamic programming. It is known that a quantized signal does not always contain signal values that can be represented with a given bit-depth. This property is called amplitude sparseness. Because quantization is the amplitude discretization of signal value, amplitude sparseness is closely related to the design of the quantizer. Since signal values with zero frequency do not affect quantization error, there is the potential to reduce complexity when designing the optimum quantizer by skipping the processing of signal values that have zero frequency. However, conventional methods on DP quantization do not design for amplitude sparseness and so are unduly complex. In this paper, we propose an algorithm that yields an optimum quantizer that minimizes quantization error with reduced complexity given the existence of amplitude sparseness.

Index Terms— quantization, sparseness, dynamic programming

1. INTRODUCTION

The purpose of quantization [1] is to generate quantization indices based on a given metric. If the metric of quantization involves distortion (quantization error) caused by the quantization process, the design of the optimal quantizer leads to a kind of minimization problem, that is, we should generate quantization indices that can minimize the quantization error. A typical quantization error expression is the sum of square error (SSE). Quantization schemes are classified into two types: conversion from a continuous signal into a discrete one, and conversion from finely discrete signal to coarser discrete one. This manuscript focuses on the latter type. The latter type is common in bit depth conversion, and is required for display adaptation [2] [3], bit-depth scalable coding [4] [5] and HDR video coding [**?**].

There are two approaches to solve the above-mentioned minimization problem: analytical optimization, which calculates optimal solutions analytically, and numerical optimization, which computes optimal solutions based on numerical computation. If the probability density function (PDF) of quantized data can be represented in some particular parametric forms, for example, uniform distribution, Gaussian distribution or Laplace distribution, you can adopt analytical optimization which analytically optimizes the quantization indices for symbols generated from these PDFs. However, such analytical optimization approaches are seldom used because the PDF of quantized data generally can not be represented in such parametric forms.

Thus numerical optimization approaches are more common as they do not require any particular parametric form of the PDF. Representative one is Lloyd-Max quantization algorithm (LM quantization) [6] [7]. However, LM quantization can not guarantee optimal solutions. In order to design an optimal quantizer, adaptive quantization algorithms based on dynamic programming (DP quantization) are studied [8]. As low complexity algorithms for designing DP quantizers, [9] address the minimization of the quantization error subject to a convexity constraint, while [10] uses matrix search to find optimal solutions for DP quantization.

When we design an optimal quantizer for an image signal, it is important to note most image signals exhibit amplitude sparseness of signal values, that is, some pixel values are never used or used very infrequently. An image with amplitude sparseness does not contain all signal values that can be represented by the given bit-depth. For example, if an image whose bit-depth is 10 bits has amplitude sparseness, its image contains fewer than 1024 different signal values, even though the image can represent up to 1024 signal values. Some studies[11] [12] [13] on image coding report that they can improve coding efficiency by considering amplitude sparseness. Thereafter, unless otherwise specified, amplitude sparseness is referred to as sparseness.

A histogram of an image with amplitude sparseness contains so insignificant elements whose frequency is zero. Signal values corresponding to insignificant elements do not af-

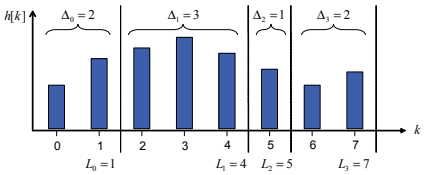

Fig. 1. Example of parameters for quantization

fect quantization error. Therefore, by designing the quantization process to properly account for insignificant elements we can reduce the complexity while minimizing quantization error. However, conventional DP quantization methods do not consider sparseness and so their efficiency has room for improvement. In this paper, we propose an algorithm that reduces the complexity of DP quantization for image signals with sparseness, while well minimizing quantization error.

2. FORMULATION OF QUANTIZER DESIGN

We formulate the design of a quantizer that translates a discrete signal with K-level to one with M-level ($M < K$). For this formulation, we use the histogram of the signal as the input of the quantizer. The k-th element of the histogram is $h[k]$ ($k = 0, \cdots, K-1$), which is the frequency of signal value k. The formulated quantizer is defined with two kinds of parameters Δ_m and L_m; Δ_m is the width of the m-th interval of the histogram. L_m is the upper boundary of the m-th interval. These parameters are described as:

$$\begin{cases} L_m = \sum_{j=0}^{m} \Delta_j - 1 \quad (m = 0, \cdots, M-2) \\ L_{M-1} = K - 1 \end{cases} \tag{1}$$

Henceforth, the m-th interval $[L_m - (\Delta_m - 1), L_m]$ of the histogram is called the m-th bin. Since each bin has at least one element, L_m $(0 \leq m \leq M-2)$ is restricted to the following range:

$$m \leq L_m \leq K - (M - m) \tag{2}$$

Fig.1 illustrates the above-mentioned parameters for a histogram with eight elements ($K = 8$) quantized into one with four bins ($M = 4$). This figure shows that the bins contain $2(= \Delta_0)$ elements, $3(= \Delta_1)$ elements, $1(= \Delta_2)$ element, and $2(= \Delta_3)$ elements of the input histogram; and the upper boundaries of the bins become $L_0 = \Delta_0 - 1 = 1$, $L_1 = L_0 + \Delta_1 = 4$, $L_2 = L_1 + \Delta_2 = 5$, and $L_3 = L_2 + \Delta_3 = 7$.

The quantizer is designed around minimizing the quantization error created by approximating all elements in the m-th bin $[L_m - (\Delta_m - 1), L_m]$ in the histogram with representative value $\hat{c}(L_m - (\Delta_m - 1), L_m)$. As the quantization error of the m-th bin $[L_m - (\Delta_m - 1), L_m]$, we use the sum of square error $e(L_m - (\Delta_m - 1), L_m)$ defined as:

$$\begin{aligned} & e(L_m - (\Delta_m - 1), L_m) \\ &= \sum_{k=L_m-\Delta_m+1}^{L_m} \{k - \hat{c}(L_m - (\Delta_m - 1), L_m)\}^2 h[k] \end{aligned} \tag{3}$$

where $\hat{c}(L_m - (\Delta_m - 1), L_m)$ is the integer value that is the closest to the centroid of the m-th bin. The centroid is defined as:

$$c(L_m - (\Delta_m - 1), L_m) = \frac{\sum_{k=L_m-(\Delta_m-1)}^{L_m} kh[k]}{\sum_{k=L_m-(\Delta_m-1)}^{L_m} h[k]} \tag{4}$$

Optimization of the quantizer means finding the parameters that minimize the following summation of quantization error

$$\begin{aligned} & (\Delta_0^*, \cdots, \Delta_{M-1}^*) \\ &= \underset{\Delta_0, \cdots, \Delta_{M-1}}{\arg\min} \{ \sum_{m=0}^{M-1} e(L_m - (\Delta_m - 1), L_m) \} \end{aligned} \tag{5}$$

3. SPARSE DP QUANTIZATION

3.1. Key point for complexity reduction

We introduce a complexity reduction algorithm for DP quantization that focuses on insignificant elements. When the frequency of signal value $L_m + 1$ is zero, that is, $h[L_m + 1] = 0$, the quantization error of interval $[L_m - (\Delta_m - 1), L_m + 1]$ in a histogram is equal to that of interval $[L_m - (\Delta_m - 1), L_m]$. This is because the addition of $h[L_m + 1](= 0)$ to the quantized interval has no effect on quantization error. Thus, when minimizing quantization error it is enough to consider only significant elements, i.e. elements whose frequencies are not zero.

In order to verify the above hypothesis on the sparseness of image signals, we measured the sparseness of standard images specified in section **4**. Sparseness is defined as the ratio of the number of insignificant elements to the number of all elements, as follows:

$$\text{Sparseness} = \frac{\text{the number of insignificant elements}}{\text{the number of all elements}} \tag{6}$$

As shown in Table 1, we confirmed that all these images exhibit some degree of sparseness.

In order to describe the proposed quantization algorithm, we define some symbols and look-up tables below. Index k, which identifies elements of histogram $h[k]$ ($k = 0, \cdots, K-1$) and index $\tilde{k}$ ($\tilde{k} = 0, \cdots, \tilde{K}-1$), which identifies the significant elements of histogram $h[k]$ are called *element index* and *significant element index*, respectively. Table $Z[k]$ ($k = 0, \cdots, K-1$) contains the number of insignificant elements belonging to interval $[0, k]$ of the histogram. Table $F[\tilde{k}]$ contains the element index corresponding to the $\tilde{k}$-th insignificant element. Table $\Psi_u[m]$ ($m = 0, \cdots, M-1$) contains the

Table 1. Sparseness of standard images (cells in the "Sparseness" column represent values output by equation (6))

Quantized signal	Sparseness [%]
Image1	7.5
Image2	8.2
Image3	11.2
Image4	19.6

maximum value of the significant index that can be the upper bound of the m-th bin. Table $\Psi_l[m]$ $(m = 0, \cdots, M-1)$ contains the minimum value of the significant index that can be the upper bound of the m-th bin. $\Psi_u[m]$ and $\Psi_l[m]$ $(m = 0, \cdots, M-1)$ are generated as follows:

$$\Psi_u[m] = \psi_u[m - M + K] - Z[\psi_u[m - M + K]] \quad (7)$$

$$\Psi_l[m] = \begin{cases} \psi_l[m] - Z[\psi_l[m]] & (m = 0, \cdots, M-2) \\ K - 1 - Z[K-1] & (m = M-1) \end{cases} \quad (8)$$

where, $\psi_u[m - M + K]$ is the maximum element index for significant elements in a range not greater than $m - M + K$.

$\psi_l[m]$ is the minimum element index for significant elements in a range not lower than m.

3.2. Optimal quantizer design that considers histogram sparseness

We describe the proposed algorithm of sparse DP quantization; it retains optimality (minimized quantization error) while reducing the complexity by skipping the processing of insignificant elements.

Assume we divide histogram interval $[0, F[\tilde{L}_m]]$ into $m+1$ bins. $F[\tilde{L}_m]$ indicates the $\tilde{L}_m$-th significant element. The sub-interval $[F[\tilde{L}_i - (\tilde{\Delta}_i - 1)], F[\tilde{L}_i]]$ $(i = 0, \cdots, m)$ of interval $[0, F[\tilde{L}_m]]$ is the i-th bin. $\tilde{L}_i$ is the significant element index of the upper bound of the i-th bin, and $\tilde{\Delta}_i$ is the number of significant elements in the i-th bin. We compute quantization error $e(F[\tilde{L}_i - (\tilde{\Delta}_i - 1)], F[\tilde{L}_i])$ created by approximating all elements in the i-th bin with a centroid value, and then store the quantization error in look up table $E[\tilde{L}_i - (\tilde{\Delta}_i - 1), \tilde{L}_i]$. We use $\tilde{S}_m[\tilde{L}_m]$ as the look up table that stores the minimum summation of quantization error $\sum_{i=0}^{m} E[\tilde{L}_i - (\tilde{\Delta}_i - 1), \tilde{L}_i]$. The minimum value of $\sum_{i=0}^{m} E[\tilde{L}_i - (\tilde{\Delta}_i - 1), \tilde{L}_i]$ is achieved with the optimal set of $\tilde{\Delta}_m, \cdots, \tilde{\Delta}_0$. Note that $\tilde{S}_m[\tilde{L}_m]$ is equal to $S_m[F[\tilde{L}_m]]$.

Since $E[\tilde{L}_m - (\tilde{\Delta}_m - 1), \tilde{L}_m]$ depends on significant index $\tilde{L}_m$ of the upper bound of the m-th bin and the number of significant elements $\tilde{\Delta}_m$ in the m-th bin, the value stored in $\tilde{S}_m[\tilde{L}_m]$ is computed from $\tilde{S}_{m-1}[\tilde{L}_m - \tilde{\Delta}_m]$ as follows:

$$\tilde{S}_m[\tilde{L}_m] = \min_{\tilde{\Delta}_m} \left[\tilde{S}_{m-1}[\tilde{L}_m - \tilde{\Delta}_m] + E[\tilde{L}_m - (\tilde{\Delta}_m - 1), \tilde{L}_m] \right] \quad (9)$$

where $m = 1, \cdots, M-1$. Using a recursive equation (9), the computation of $\tilde{S}_m[\tilde{L}_m]$ is equivalent to selecting the optimal parameters among $\tilde{\Delta}_m = 1, \cdots, \tilde{L}_m - \Psi_l[m-1]$. Considering that the upper bound and the lower bound of significant indices in the m-th bin are defined as $\Psi_u[m]$ and $\Psi_l[m]$, respectively, we have $\tilde{L}_m = \Psi_l[m], \cdots, \Psi_u[m]$. The value stored in $\tilde{S}_m[\tilde{L}_m]$ is used in computing $\tilde{S}_{m+1}[\tilde{L}_{m+1}]$.

Letting $\tilde{\Delta}_m^{(\tilde{L}_m)}$ denote $\tilde{\Delta}_m$ that minimizes the right side of equation (9), we store the optimal upper bound of the $m-1$-th bin for each upper bound $\tilde{L}_m$ $(= \Psi_l[m], \cdots, \Psi_u[m])$ of the m-th bin in table defined by $T_{m-1}[\tilde{L}_m]$ as follows:

$$T_{m-1}[\tilde{L}_m] = \tilde{L}_m - \tilde{\Delta}_m^{(\tilde{L}_m)}$$

In the case of $m = 0$, $\tilde{S}_0[\tilde{L}_0]$ represents the quantization error caused by approximating histogram interval $[0, F[\tilde{L}_0]]$ with the centroid of the interval; this yields:

$$\tilde{S}_0[\tilde{L}_0] = E[0, F[\tilde{L}_0]]$$

The optimum parameters $(\Delta_0^*, \cdots, \Delta_{M-1}^*)$ are obtained through the following process. The minimization problem of equation (5) becomes:

$$\min_{\tilde{\Delta}_{M-1}} \left[\tilde{S}_{M-2}[\tilde{L}_{M-1} - \tilde{\Delta}_{M-1}] + E[\tilde{L}_{M-1} - (\tilde{\Delta}_{M-1} - 1), \tilde{L}_{M-1}] \right]$$

We express $\tilde{\Delta}_{M-1}^*$ that minimizes the above equation as follows:

$$\tilde{\Delta}_{M-1}^* = \arg \min_{\tilde{\Delta}_{M-1}} \Big[\tilde{S}_{M-2}[\tilde{L}_{M-1} - \tilde{\Delta}_{M-1}] + E[\tilde{L}_{M-1} - (\tilde{\Delta}_{M-1} - 1), \tilde{L}_{M-1}] \Big]$$

Since the only possible value of $\tilde{L}_{M-1}$ is $K - Z[K-1] - 1$, we have $\tilde{L}_{M-1} = K - Z[K-1] - 1$. Using $\tilde{L}_{M-1}$ and $\tilde{\Delta}_{M-1}^*$, the optimal significant element index of the upper bound of the $M-2$-th bin can be obtained as $\tilde{L}_{M-2}^* = \tilde{L}_{M-1} - \tilde{\Delta}_{M-1}^* = K - Z[K-1] - 1 - \tilde{\Delta}_{M-1}^*$. Since the optimal significant element index of the upper bound of the $M-3$-th bin for $\tilde{L}_{M-2}^*$ is stored in $T_{M-3}[\tilde{L}_{M-2}^*]$, let $\tilde{L}_{M-3}^* = T_{M-3}[\tilde{L}_{M-2}^*]$. Referring tables similarly, we obtain $\tilde{L}_{M-4}^* = T_{M-4}[\tilde{L}_{M-3}^*]$,$\cdots$, $\tilde{L}_0^* = T_0[\tilde{L}_1^*]$ as the significant element indices of the upper bound of each bin. By accessing $F[]$ with these obtained significant element indices $\tilde{L}_{M-1}$, $\tilde{L}_{M-2}^*$,$\cdots$, $\tilde{L}_0^*$, we have the element indices of the upper bound of each bin. As a result, the intervals of each bin are derived as follows: $\Delta_{M-1}^* = F[\tilde{L}_{M-1}] - F[\tilde{L}_{M-2}^*] = K - 1 - F[\tilde{L}_{M-2}^*]$, $\Delta_{M-2}^* = F[\tilde{L}_{M-2}^*] - F[\tilde{L}_{M-3}^*]$,$\cdots$, $\Delta_1^* = F[\tilde{L}_1^*] - F[\tilde{L}_0^*]$, $\Delta_0^* = F[\tilde{L}_0^*] + 1$.

4. EXPERIMENTS

We performed the following experiments to investigate the effectiveness of our quantization algorithm from the viewpoint of complexity.

(a) Image1 (b) Image2

(b) Image3 (d) Image4

Fig. 2. Thumbnail images of input signals that were quantized

As the input signal, we used the sequences in *ITE/ARIB Hi-Vision Test Sequence 2nd Edition*. The sequences have progressive scan format with resolution of 1920 $\times$ 1080 pixels/frame. Luminance signal (10 bit scale) of the first frame of each sequence was used in the following evaluation experiments. The bit-depth of input signal means $K = 1024$. In these experiments, we set $M = 128, 256$ as the number of bins. These experiments were performed on a computer with CPU:Intel core i5 (2.6GHz) and memory:8GB.

We compared sparse DP quantization (abbreviated to SDP-Q in the following tables and figures) with DP quantization (abbreviated to DP-Q hereafter) in order to evaluate complexity reduction achieved by sparse DP quantization. DP quantization (abbreviated to DP-Q in the following tables and figures) The results are shown in Fig. 2, where processing time is the average processing time of 1000 trials. We used the following metric to evaluate the complexity reduction attained by sparse DP quantization:

$$\text{complexity reduction ratio} = \frac{\text{processing time of DP-Q} - \text{processing time of SDP-Q}}{\text{processing time of DP-Q}} \quad (10)$$

In order to elucidate the overall algorithm attributes for all sequences, Table 2 shows average processing time of DP-Q and SDP-Q for all sequences at all M values. From this table, we can confirm that sparse DP quantization can reduce complexity by 21.2 to 24.2% on average over DP quantization.

The results show that the complexity reduction ratio increases as M decreases. The reason for this is as follows. When the number of insignificant elements M is a constant, the ratio of the number of insignificant elements to the number of bins increases as the number of bins decreases. Sparse DP quantization achieves complexity reduction by skipping the processing of insignificant elements. Thus, it is understandable that complexity reduction ratio (sparse DP quantization to DP quantization) becomes large, as M decreases.

Let us consider the complexity reduction of sparse DP quantization from the viewpoint of sparseness. As shown in Table 1 in **3.1**, image sparseness increases in the order of

Table 2. Processing time of DP-Q and SDP-Q (cells in the "Reduction ratio" column represent values defined in equation (10))

(a) M=128

M	DP-Q [msec]	SDP-Q [msec]	reduction ratio [%]
image1	226	193	16.2
image2	228	189	17.1
image3	225	176	21.8
image4	223	130	41.7
average	226.0	171.5	24.2

(b) M=256

M	DP-Q [msec]	SDP-Q [msec]	reduction ratio [%]
image1	239	201	15.9
image2	239	199	16.7
image3	238	191	19.7
image4	235	159	32.3
average	237.8	187.5	21.2

"image1", "image2", "image3", and "image4". According to Table 2, we can confirm that the complexity reduction ratio improves as sparseness increases.

5. CONCLUSIONS

This paper tackled complexity reduction for dynamic programming (DP) quantization by focusing on the sparseness of signal values. The proposed method, called sparse DP quantization, keeps the optimality of DP quantization in terms of minimizing the quantization error. Specifically, sparse DP quantization can reduce the complexity of DP quantization without increasing quantization error. Experiments on standard images showed that sparse DP quantization offers, on average, 21.2 to 24.2% less complexity than DP quantization.

Sparse DP quantization can be used as a technology to complement conventional methods [9] [10] since the conventional methods take approaches that do not depend on the sparseness of signal values. Therefore, by combining sparse DP quantization and conventional methods, the complexity of DP quantization can be further reduced.

6. REFERENCES

[1] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[2] E. Reinhard, S. Pattanaik, G. Ward, and P. Debevec, *High Dynamic Range Imaging: Acquisition, Display,*

and Image-Based Lighting, Morgan Kaufmann Publisher, 2005.

[3] E. François, D. Rusanovskyy, P. Yin, P. Topiwala, G. Sullivan, and M. Naccari, "Signalling, backward compatibility and display adaptation for HDR/WCG video coding, draft 1," JCTVC-Y1012, Oct. 2016.

[4] J. Boyce, Y. Ye, J. Chen, and A. Ramasubramonian, "Overview of SHVC: Scalable extensions of the high efficiency video coding standard," *IEEE Trans. Circuits Syst. Video*, vol. 26, no. 1, pp. 20–34, 2015.

[5] *ISO/IEC 18477-2:2016: Information technology: Scalable compression and coding of continuous-tone still images – Part 2: Coding of high dynamic range images*, 2016.

[6] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, pp. 129–136, Mar. 1982.

[7] J. Max, "Quantizing for minimum distortion," *IRE. Trans. Inf. Theory*, vol. IT-7, pp. 7–12, Mar. 1960.

[8] J. D. Bruce, *Optimum quantizer*, Ph.D. thesis, M.I.T., May 1964.

[9] D. Sharma, "Design of absolutely optimal quantizers for a wide class of distortion measures," *IEEE Trans. Inf. Theory*, vol. 24, no. 6, pp. 693–702, Nov. 1978.

[10] X. Wu, "Optimal quantization by matrix searching," *Journal of Algorithms*, vol. 12, no. 4, pp. 663–673, Dec. 1991.

[11] P. Ferreira and A. J. Pinho, "Why does histgram packing improve lossless compression rates ?," *IEEE Signal processing letters*, vol. 9, no. 8, pp. 259–261, 2002.

[12] M. Aguzzi and M. Albanesi, "A novel approach to sparse histogram image lossless compression using JPEG 2000," *Electronic Letters on Computer Vision and Image Analysis*, vol. 5, no. 4, pp. 24–46, 2006.

[13] E. Nasr-Esfahani, S. Samavi, N. Karimi, and S. Shiran, "Near lossless image compression by local packing of histogram," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 1197–1200, 2008.