IEICE TRANS. FUNDAMENTALS, VOL.E103–A, NO.12 DECEMBER 2020

PAPER Special Section on Smart Multimedia & Communication Systems

Multi-Layered DP Quantization Algorithm

Yukihiro BANDOH[†a)], **Seishi TAKAMURA**[†], *and* **Hideaki KIMATA**[†], ***Senior Members***

SUMMARY Designing an optimum quantizer can be treated as the optimization problem of finding the quantization indices that minimize the quantization error. One solution to the optimization problem, DP quantization, is based on dynamic programming. Some applications, such as bit-depth scalable codec and tone mapping, require the construction of multiple quantizers with different quantization levels, for example, from 12bit/channel to 10bit/channel and 8bit/channel. Unfortunately, the above mentioned DP quantization optimizes the quantizer for just one quantization level. That is, it is unable to simultaneously optimize multiple quantizers. Therefore, when DP quantization is used to design multiple quantizers, there are many redundant computations in the optimization process. This paper proposes an extended DP quantization with a complexity reduction algorithm for the optimal design of multiple quantizers. Experiments show that the proposed algorithm reduces complexity by 20.8%, on average, compared to conventional DP quantization.

key words: *quantization, dynamic programming, bit-depth scalability, multi-layered structure*

1. Introduction

Quantization [1] is a scheme that attempts to replicate discrete signals by generating quantization indices based on a given metric. If the metric of quantization permits distortion (quantization error) to occur within the quantization process, designing an optimal quantizer is equivalent to a minimization problem, where the generation of quantization indices has the goal of minimizing the quantization error. A typical quantization error expression is the sum of square error (SSE). Quantization schemes are classified into two types from the viewpoint of the input signal: the first type converts a continuous signal into a discrete one, while the second type converts a higher resolution discrete signal into a coarse discrete signal. This manuscript focuses on the latter type, and so assumes the input of high resolution discrete signals. This functionality is common in bit-depth conversion, and is required for display adaptation [2]–[4], bit-depth scalable coding [5]–[7] and HDR video coding [8]–[11].

Two approaches are used to solve the above-mentioned minimization problem: analytical optimization, which calculates optimal solutions analytically, and numerical optimization, which identifies optimal solutions based on numerical computation. Analytical optimization is limited to the case that the probability density function (PDF) of quantized data can be represented in some particular parametric forms, such as Gaussian distribution. By contrast, numerical optimization methods are more common as they do not require the PDF to follow any particular parametric form. A representative method is the Lloyd-Max quantization algorithm (LM quantization) [12], [13]. Unfortunately, LM quantization cannot guarantee optimal solutions. For designing optimal quantizers, adaptive quantization algorithms based on dynamic programming (DP quantization), which originated in Bruce's algorithm [14], have been studied. For example, so as to reduce the complexity of DP quantization, Sharma [15] attempts to minimize the quantization error subject to a convexity constraint, Wu [16] uses matrix search to find optimal solutions for DP quantization, and Bandoh [17] focuses on the amplitude sparseness of signal values.

Some applications for HDR images such as SDR display adaptation and bit-depth scalable coding need to convert a source signal with high bit-depth into multiple formats with lower bit-depths to support a wide variety of user environments. For example, while the source signal may have been captured with 12 bits/component (HDR image), some users can observe the signal only as pseudo-HDR images because their legacy displays support only 10 bits/component or 8 bits/component. The formats needed, including bit-depth, depend on the variety of user environments. This triggered the release of bit-depth scalable codecs such as HEVC/H.265 scalability extension (SHVC) [18] and AVC/H.264 scalability extension (SVC) [19] to support a multi-layered structure that includes a base-layer (8 bit/component) and multiple enhancement layers (10 and 12 bit/component). The result is a processed bitstream that realizes SDR display adaptation.

Our target is the optimal design of a multi-level quantizer that supports multiple quantization levels. A straightforward approach is to apply DP quantization to each quantization level, simultaneously. This approach treats the quantizers of each specific level independently. It does not consider redundant computations between the quantizers with different quantization levels. The above mentioned studies on complexity reduction of DP quantization follow this approach category. To maximize the complexity reduction when optimizing a multi-level quantizer, we focus on eliminating the redundant computations inherent in processing different quantization levels, while minimizing quantization error.

In this paper, we propose an algorithm that reduces the complexity of DP quantization for multiple quantization lev-

Manuscript received February 7, 2020.
Manuscript revised April 23, 2020.

[†]The authors are with NTT Media Interigence Laboratories, NTT Corporation, Yokosuka-shi, 239-0847 Japan.
a) E-mail: yukihiro.bandoh@m.ieice.org

DOI: 10.1587/transfun.2020SMP0028

els, while well minimizing quantization error. This paper enhances the basic study of [20] with regard to three points. First, this paper presents a complete algorithm that reduces the complexity of DP quantization for multiple quantization levels, while still minimizing quantization error. Second, this paper introduces more extensive experimental results through evaluations on more kinds of image contents. Finally, this paper provides an analytical evaluation to assess the complexity of the proposed algorithm.

This paper studies a layered design for multi-level quantization as an extension of DP quantization and differs from conventional studies on DP-based layered quantization, as described below. Muresan [21] uses DP for designing multi resolution scalar quantization (MRSQ) that generates multi-level quantizers under a restriction that requires successive refinement of partitions. Its restriction requires that a refined quantization have quantization bins that are contained within those of coarse quantization. By contrast, the proposed method in this paper assumes no such structural restriction. The proposed method challenge optimal quantization designs to support wider ranges than MRSQ. Khandani [22] studies the design of entropy constraint vector quantization based on DP. The study utilizes DP for hierarchically constructing elements of a multi-dimensional vector for an efficient dictionary. Note that this study takes the position of reducing complexity at the expense of increase in quantization error. Thus, its constructed quantizer does not guarantee the global optimal design.

This paper is organized as follows. Section 2.1 formulates the problem of quantizer optimization. Section 2.2 introduces a DP quantization that is the key component of our proposed algorithm. Section 2.3 elucidates optimal quantizer design from the viewpoint of path search on a trellis diagram. Section 2.4 enhances DP quantization for designing multi-level quantizers in terms of dropping duplicate computations on adjacent quantization levels. As reference information, notations used in Sect. 2 are summarized in Table 1. Section 3 details the results of experiments conducted on the proposed method. Finally, Sect. 4 presents our conclusions.

2. Multi-Level Quantizer Design

2.1 Formulation of M-Level Quantizer Design

We formulate the design of a quantizer that translates a K-level discrete signal into an M-level equivalent ($M < K$). We realize this by processing a histogram of the signal as the input to 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, which is called the M-level quantizer, is defined using parameters Δ_m and L_m; Δ_m is the length of the m-th sub-interval of the histogram. L_m is the upper boundary of the m-th sub-interval in the histogram. The boundaries are described below starting with the parameters given by:

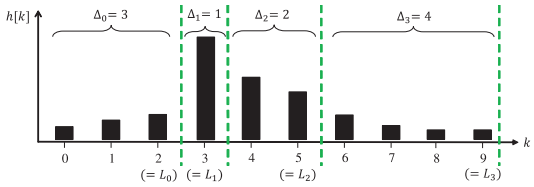


Fig. 1 Example of parameters used in quantization.

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

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 ($0 \le m \le M-2$) is restricted to the following range:

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

Figure 1 illustrates the above-mentioned parameters for a histogram with ten elements ($K = 10$) quantized into four bins ($M = 4$). This figure shows that the bins contain $3(= \Delta_0)$ elements, $1(= \Delta_1)$ element, $2(= \Delta_2)$ elements, and $4(= \Delta_3)$ elements of the input histogram, and the upper boundaries of the bins are $L_0 = \Delta_0 - 1 = 2$, $L_1 = \Delta_0 + \Delta_1 - 1 = 3$, $L_2 = \Delta_0 + \Delta_1 + \Delta_2 - 1 = 5$, and $L_3 = \Delta_0 + \Delta_1 + \Delta_2 + \Delta_3 - 1 = 9$.

Quantizer design is based on minimizing the quantization error created by approximating all values in the m-th bin $[L_m - (\Delta_m - 1), L_m]$ of the histogram with representative value $\hat{c}(\Delta_m, L_m)$. To assess the quantization error of the m-th bin, we use the summation of square error $e(\Delta_m, L_m)$ defined as follows:

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

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

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

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

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

2.2 M-Level Quantizer Design Based on Dynamic Programing

Given that the quantization error $e(\Delta_m, L_m)$ of the m-th bin depends on the boundary, L_m, of the m-th bin and the length, Δ_m, of the same bin, dynamic programming based approaches (DP quantization) have been used to solve the optimization problem represented by Eq. (5).

Table 1 Notations.

Symbol	Description
K	the number of levels of input signal
M_1,M_2	the number of levels of quantized signal, $M_1 < M_2 < K$
$h[k]$	the k-th element ($k = 0, \cdots, K-1$) of the histogram of input signal, which is abbreviated as "the histogram" in this table
L_m	the upper boundary of the m-th interval in the histogram ($m = 0, \cdots, M-1$)
Δ_m	the length of the m-th interval of the histogram ($m = 0, \cdots, M-1$)
$e(\Delta_m, L_m)$	quantization error of the m-th interval $[L_m - (\Delta_m - 1), L_m]$ of the histogram
$E[\Delta_m, L_m]$	the (Δ_m, L_m)-th element of a look-up table for referring the value of $e(\Delta_m, L_m)$
$S_m[L_m]$	the minimum summation of quantization error if the interval $[0, L_m]$ of the histogram is divided into $m+1$ sub-intervals
$T_{m-1}[L_m]$	the optimal boundary of the $m-1$-th bin which is next to the m-th bin with boundary L_m

[1)] *Element index* is an index to identify each element of the histogram.

DP quantization focuses on the recurrence relation of quantization error. We define $S_m[L_m]$ for each L_m ($m = 0, \cdots, M-1$) as the minimum summation of quantization error $\sum_{i=0}^{m} e(\Delta_i, L_i)$ where interval $[0, L_m]$ of histogram $h[k]$ ($k = 0, \cdots, K-1$) is divided into $m+1$ bins. Since $e(\Delta_m, L_m)$ depends on L_m and Δ_m, $S_m[L_m]$ can be expressed using $S_{m-1}[L_m - \Delta_m]$ given by the following recursive equation:

$$S_m[L_m] = \min_{\Delta_m} \{S_{m-1}[L_m - \Delta_m] + e(\Delta_m, L_m)\} \tag{6}$$

where $m = 1, \cdots, M-1$ and $L_m = m, \cdots, K-(M-m)$. Equation (6) says that computing $S_m[L_m]$ results in the selection of the best parameter among the values of $\Delta_m = 1, \cdots, L_m - m + 1$. Solving Eq. (6) from $m = 1$ up to $m = M-1$ recursively, the minimization problem of Eq. (5) can be written as follows:

$$\min_{\Delta_{M-1}} \{S_{M-2}[L_{M-1} - \Delta_{M-1}] + e(\Delta_{M-1}, L_{M-1})\} \tag{7}$$

2.3 Interpretation of Optimal Quantizer on Trellis Transition Diagram

We use a trellis transition diagram to better explain the optimization process of DP quantization. This interpretation will be useful in understanding the proposed algorithm described in Sect. 2.4. The trellis transition diagram of Fig. 2 illustrates the quantization result for the example shown in Fig. 1. The traversed path, indicated by the bold blue lines in Fig. 2, corresponds to the quantization result shown in Fig. 1. In Figd. 2, the vertical axis and the horizontal axis represent signal values $k \in \{0, 1, \cdots, 9\}$ to be quantized and quantization indices $m \in \{0, 1, 2, 3\}$, respectively. The trellis transition diagram consists of nodes and paths. The node at (k, m) has a state wherein the interval $[0, k]$ in the histogram is approximated with $m+1$ levels, and the upper bound of the $m+1$-th bin is the k-th element of the histogram. The path from node $(k-\Delta, m-1)$ to (k, m) has quantization error of the m-th bin corresponding to interval $[k-\Delta+1, k]$. Thus, the design of the optimal quantizer that translates a K-level discrete signal into an M-level one can be represented as optimal path search from the node at $(-1, -1)$ to that at $(K-1, M-1)$ over the trellis transition diagram. Note that the node at $(-1, -1)$ is a dummy node introduced as the start node and does not have the above-mentioned state.

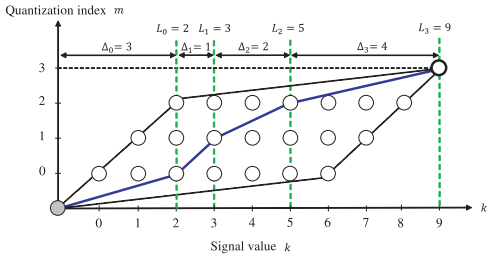


Fig. 2 Traversed path corresponding to quantization shown in Fig. 1.

The transitions over the diagram need to satisfy the following constraints:

(A) move one step in the upward direction per transition
(B) move at least one step in the right direction per transition

Considering the above constraints, in the case of the quantization example ($K = 10, M = 4$) shown in Fig. 1, nodes that can be transited in the trellis transition diagram are restricted to those within the region delineated by broken-lines as shown in Fig. 2. This is generalized as follows: the nodes on the m-th row ($m = 0, \cdots, M-2$) exist in the range of $k = m, \cdots, K-M+m$.

2.4 Complexity Reduction for Designing Optimal Multi-Level Quantizers

We now turn to the design of the optimization of multiple quantizers that support multiple quantization levels. Our example is the optimization of an M_1-level quantizer and an M_2-level quantizer that both have as their inputs the same K-level signal ($M_1 < M_2 < K$). Note that although the following discussion deals with just two quantization levels, the same discussion applies to the general case of more than two levels without loss of generality. A straightforward approach to the design is to apply DP quantization to each quantization level independently. Hereinafter, this approach is referred to as *simultaneous DP quantization*. Simultaneous DP quantization focuses on complexity reduction for a specific level of quantization. However, it does not address redundant computations between quantizers with different quantization levels. To further reduce the complexity of multi-level quantizers, we focus on eliminating the redun-

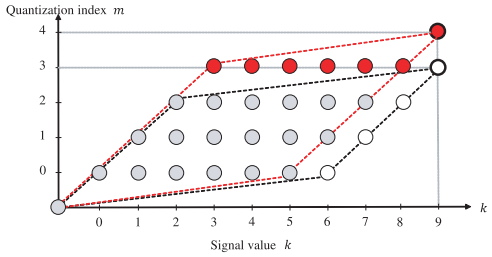


Fig. 3 Trellis transition diagram for the example of $M = 4$ and $M = 5$.

dant computations permitted by simultaneous DP quantization.

We introduce the basic concept of our proposal using the example shown in Fig. 3. In detail, we discuss a multi-level quantizer that quantizes a 10-level signal ($K = 10$) to a 4-level signal ($M_1 = 4$) and a 5-level signal ($M_2 = 5$). Figure 3 illustrates two trellis transition diagrams: one for 4-level quantization and one for 5-level. For the case of $M_1 = 4$, traversed nodes are restricted to the white and gray nodes within the parallelogram area bounded by the black broken-lines. For the case of $M_2 = 5$, traversed nodes are restricted to the red and gray nodes within the parallelogram area surrounded by the red broken-lines. The accumulated cost incurred by the gray nodes is the same for both $M_1 = 4$ and $M_2 = 5$. When we try to find the optimal quantizer for $M_2 = 5$ after obtaining the optimal one for $M_1 = 4$, it is not necessary to recompute the gray nodes for the case of $M_2 = 5$. This is because those gray nodes are identical to those determined for $M_1 = 4$. For the case of $M_2 = 5$, it is sufficient to compute just the red nodes. This suppresses duplicate computation, which reduces complexity while retaining solution optimality.

Generalizing the above example, let us consider the hierarchical optimization of an M_1-level quantizer and an M_2-level quantizer whose inputs are the same K-level signal. The hierarchical optimization is designed assuming an M_1-level quantizer that satisfies

$$(\Delta_0^{(1)^*}, \cdots, \Delta_{M_1-1}^{(1)^*}) = \underset{\Delta_0,\cdots,\Delta_{M_1-1}}{\arg\min} \left\{ \sum_{m=0}^{M_1-1} e(\Delta_m, L_m) \right\} \tag{8}$$

which is followed by an M_2-level quantizer that satisfies

$$(\Delta_0^{(2)^*}, \cdots, \Delta_{M_2-1}^{(2)^*}) = \underset{\Delta_0,\cdots,\Delta_{M_2-1}}{\arg\min} \left\{ \sum_{m=0}^{M_2-1} e(\Delta_m, L_m) \right\} \tag{9}$$

The M_1-level quantizer is generated by solving Eq. (8) with DP quantization. The M_2-level quantizer is generated by solving Eq. (9) with our improved DP quantization proposal that eliminates wasteful re-computation by reusing some of the results of DP quantization for the M_1-level quantizer. Basically, the M_2-level quantizer incurs accumulated costs of $S_m[k]$ ($m = 0, \cdots, M_2-2, k = m, \cdots, K-M+m$) based on Eq. (6) in ascending order of m. However, it is different from the M_1-level quantizer in that not all $S_m[k]$ are computed from scratch. Since $S_m[k]$ in the range of $0 \le m \le M_1 - 2$ are the same as those for the M_1-level quantizer, they can be obtained by simply reusing the values computed for the M_1-level quantizer. What needs to be computed from scratch for the M_2-level quantizer is limited to $S_m[k]$ in the range of $M_1 - 1 \le m \le M_2 - 1$. The above approach is referred to as *layered DP quantization*.

Figure 4 shows the pseudo-code of the multi-level quantizer stated above. Instruction 2 in Fig. 4 generates a look-up-table that stores the quantization error of every interval in the histogram. This look-up-table is shared between M_1-level quantizer and M_2-level quantizer. Instructions 3 to 15 correspond to the design of a M_1-level quantizer using the DP quantization algorithm. In these processes, instructions 3 to 8 solve the minimization problem recursively based on Eq. (6), as described in Sect. 2.2. Instruction 9 stores the optimal boundary of the $m-1$-th bin L_m in table $T_{m-1}[L_m]$ for later reference. Instructions 10 to 14 obtain the optimal parameters $(\Delta_0^{(1)^*}, \cdots, \Delta_{M-1}^{(1)^*})$ from the following process, which is called the back-track process. Since the possible value of L_{M-1} is limited to $K-1$, as the optimal value of L_{M-1}, we have $L_{M-1}^{(1)^*} = K-1$. By using $L_{M-1}^{(1)^*} = K-1$ as a start point, and referring to table $T_m[]$, we obtain $L_{M-2}^* = T_{M-2}[L_{M-1}^*], \cdots, L_0^* = T_0[L_1^*]$. Using these values $L_{M-1}^{(1)^*}, \cdots, L_0^{(1)^*}$, we derive $\Delta_{M-1}^{(1)^*} = L_{M-1}^{(1)^*} - L_{M-2}^{(1)^*}, \cdots, \Delta_1^{(1)^*} = L_1^{(1)^*} - L_0^{(1)^*}, \Delta_0^{(1)^*} = L_0^{(1)^*} + 1$ for M_1-level quantizer.

Instructions 16 to 27 design the M_2-level quantizer based on the algorithm provided in Sect. 2.4. Instruction 16 loads accumulated costs which are computed in generating the M_1-level quantizer. By reusing these accumulated costs, instructions 17 to 21 focus on $m = M_1 - 1, \cdots, M_2 - 1$, i.e. the nodes in the restricted range among the $M_1 - 1$-th row and the $M_2 - 1$-th row of the trellis transition diagram. Thus, we can eliminate the computations associated with $m = 0, \cdots, M_1 - 2$. Finally, the back-track process in instructions 21 to 27 generates the optimal parameters $(\Delta_0^{(2)^*}, \cdots, \Delta_{M-1}^{(2)^*})$ for M_2-level quantizer.

3. Experiments

We performed the following experiments in order to investigate the effectiveness of our quantization algorithm from the viewpoint of complexity reduction. As the input signal of each quantization algorithm, we used the sequences in *ITE/ARIB Ultra-high definition/wide-color-gamut standard test sequences - Series A and Series B* [23], [24]. The sequences employ the progressive scan format with resolution of 3840×2160 pixels/frame in the RGB4:4:4 color format defined in ITU-R Recommendation BT.2020. These signals were sampled at 12 bit scale, so $K = 4096$. Green channel signals of the head frame of each sequence were used in the following evaluation experiments. Given the existence of legacy displays, it is often necessary to convert high bit-depth signals into low bit-depth signals that have just ten or eight bits/channel. Accordingly, we set $M = 1024, 256$ as the number of bins. These experiments were performed on a computer with an Intel core i7 CPU (2.8GHz) and 8GB of

Load histogram $h[k]$ $(k = 0, \cdots, K-1)$ of the signal values
Load the look-up-table $E[i_e - i_s + 1, i_e]$ ($i_s \le i_e$, $i_s = 0, \cdots, K-1$, $i_e = 0, \cdots, K-1$) for quantization error of each interval in histogram $h[k]$
for $j = 0, \cdots, K - M_1$
    $S_0[j] \leftarrow E[0, j]$ */* for nodes in the 0-th row of the trellis transition diagram */*
for $m = 1, \cdots, M_1 - 1$ */* for nodes in the 1-st row to $M_1 - 1$-th row of the trellis transition diagram */*
    for $L_m = m, \cdots, K - (M_1 - m)$
        $S_m[L_m] \leftarrow \min_{\Delta_m = 1, \cdots, L_m - m + 1} \{S_{m-1}[L_m - \Delta_m] + E[L_m - (\Delta_m - 1), L_m, \lambda]\}$
        $\Delta_m^{(L_m)} \leftarrow \arg \min_{\Delta_m = 1, \cdots, L_m - m + 1} \{S_{m-1}[L_m - \Delta_m] + E[L_m - (\Delta_m - 1), L_m, \lambda]\}$
        $T_{m-1}[L_m] \leftarrow L_m - \Delta_m^{(L_m)}$
$L_{M_1-1}^{(1)^*} \leftarrow K - 1$
for $m = M_1 - 1, \cdots, 1$
    $L_{m-1}^{(1)^*} \leftarrow T_{m-1}[L_m^{(1)^*}]$
    $\Delta_m^{(1)^*} \leftarrow L_m^{(1)^*} - L_{m-1}^{(1)^*}$
$\Delta_0^{(1)^*} \leftarrow L_0^{(1)^*} + 1$
Output $\Delta_m^{(1)^*}$ for $m = M_1 - 1, \cdots, 0$ as optimum parameters of M_1-level quantizer */*Up to here : design of M_1-level quantizer */*
Load accumulated cost $S_m[j]$ $(m = 0, \cdots, M_1 - 2,\ j = m, \cdots, K - (M_1 - m))$ using DPQ with M_1-level
for $m = M_1 - 1, \cdots, M_2 - 1$ */* for nodes in the $M_1 - 1$-th row to $M_2 - 1$-th row of the trellis transition diagram */*
    for $L_m = m, \cdots, K - (M_2 - m)$
        $S_m[L_m] \leftarrow \min_{\Delta_m = 1, \cdots, L_m - m + 1} \{S_{m-1}[L_m - \Delta_m] + E[L_m - (\Delta_m - 1), L_m, \lambda]\}$
        $\Delta_m^{(L_m)} \leftarrow \arg \min_{\Delta_m = 1, \cdots, L_m - m + 1} \{S_{m-1}[L_m - \Delta_m] + E[L_m - (\Delta_m - 1), L_m, \lambda]\}$
        $T_{m-1}[L_m] \leftarrow L_m - \Delta_m^{(L_m)}$
$L_{M_2-1}^{(2)^*} \leftarrow K - 1$
for $m = M_2 - 1, \cdots, 1$
    $L_{m-1}^{(2)^*} \leftarrow T_{m-1}[L_m^{(2)^*}]$
    $\Delta_m^{(2)^*} \leftarrow L_m^{(2)^*} - L_{m-1}^{(2)^*}$
$\Delta_0^{(2)^*} \leftarrow L_0^{(2)^*} + 1$
Output $\Delta_m^{(2)^*}$ for $m = M_2 - 1, \cdots, 0$ as optimum parameters of M_2-level quantizer

Fig. 4 Optimal design of multi-level quantizer with M_1/M_2-level using multi-layered DP quantization.

memory.

In order to evaluate the complexity reduction achieved by layered DP quantization (abbreviated as LyDP-Q), we compared LyDP-Q with simultaneous DP-quantization (abbreviated as SmDP-Q) in terms of running time. As we described in 2.4, SmDP-Q designs two optimal quantizers for $M = 1024$ and $M = 256$ by applying DP quantization to each quantization level $M = 1024, 256$ independently. By contrast, LyDP-Q supports the elimination of duplicate computations between different quantization levels. For this comparison, we calculated the following metric:

$$\text{complexity reduction ratio} = \frac{\text{running time of SmDP-Q} - \text{running time of LyDP-Q}}{\text{running time of SmDP-Q}} \quad (10)$$

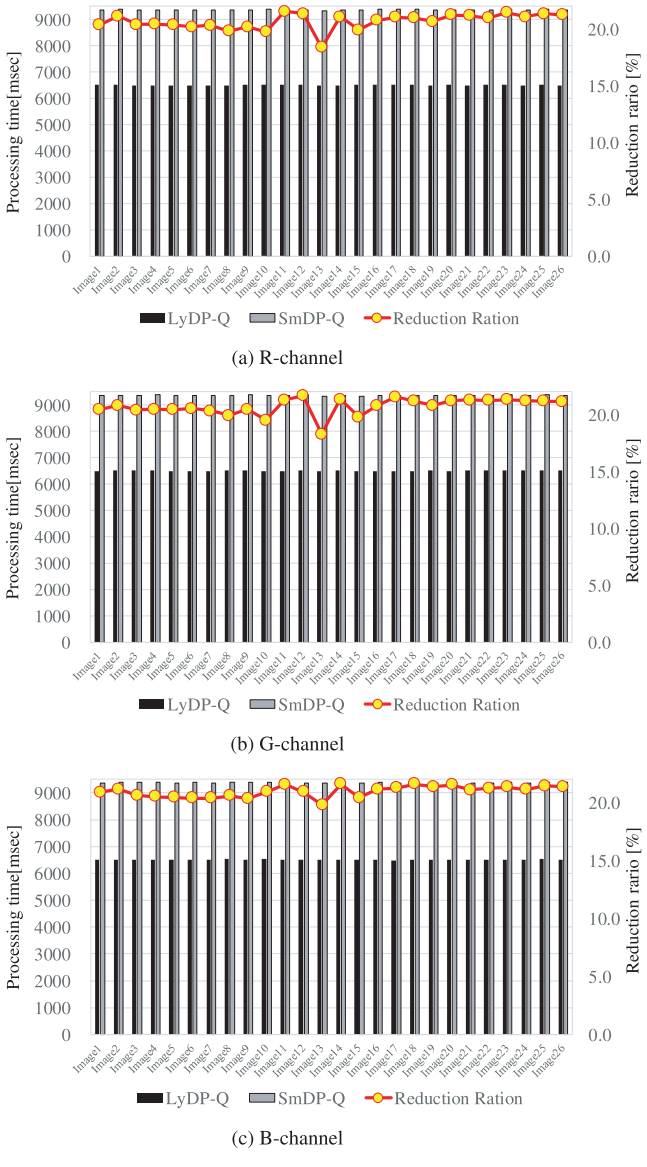
The results are shown as bar graphs in Fig. 5, where gray bars and black bars represent the running time of SmDP-Q and LyDP-Q, respectively. We confirmed that LyDP-Q yielded a 20.8% (on average) complexity reduction over SmDP-Q. The proposed algorithm leads to computation efficiency while keeping the optimality of multi-level quantization in terms of minimizing the total amount of quantization error.

Furthermore, our layered approach can be incorporated into sparse DP quantization (SDP-Q) [17] in the same way as DP quantization. Figure 6 shows the running time of our layered approach built on SDP-Q (abbreviated as LySDP-Q) and the simultaneous approach built on SDP-Q (abbreviated as SmSDP-Q). We confirmed that LySDP-Q had less complexity, 16.1% on average, than SmSDP-Q. SDP-Q pays attention to the amplitude sparseness of signal values for complexity reduction, whereas our layered approach focuses on inter-level redundancy. Therefore, as the above results show, our layered approach well complements the conventional complexity reduction scheme for DP quantization.

The proposed method with its layered strategy achieved its complexity reduction by eliminating the following computations:

- **(C1)** duplicated processes to search for optimal paths for M_2-level quantization
- **(C2)** duplicated generation of a look-up table $E[]$ for M_2-level quantization

Further details on the complexity reduction are to be found in Table 2. Table 2(a) shows average processing time of SmDP-Q and LyDP-Q for all images. The figures in column "reduction ratio" in Table 2(a) were computed by applying the same concept as Eq. (10) to the candidate paths of SmDP-Q and LyDP-Q. As a breakdown of the processing time of both algorithms, Tables 2(b) and (c) give the processing time for searching optimal path and that for generating look-up tables, respectively. As a countermeasure to the above-mentioned "C1", LyDP-Q reused a part of the computations performed for M_1-level quantization for opti-

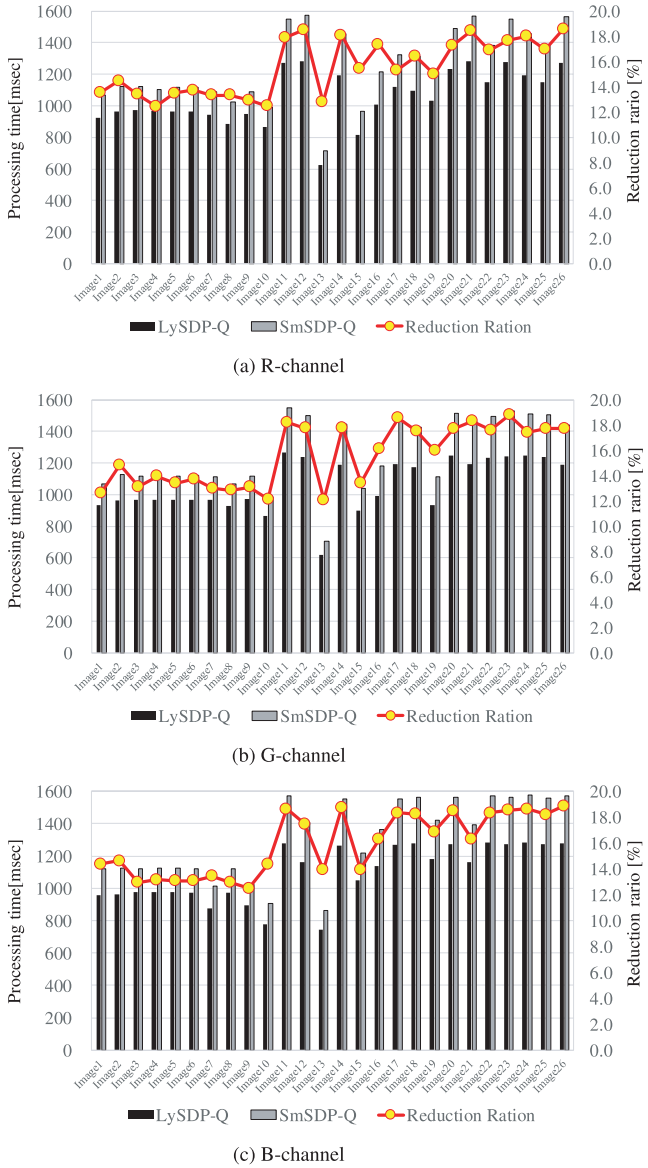


(a) R-channel

(b) G-channel

(c) B-channel

Fig. 5 Running time of LyDP-Q and SmDP-Q ("reduction ratio" is defined in (10)).



(a) R-channel

(b) G-channel

(c) B-channel

Fig. 6 Running time of LySDP-Q and SmSDP-Q ("reduction ratio" is defined by Eq. (10)).

mizing M_2-level quantization. As a result, LyDP-Q reduced the processing time of this process by 16.7% over SmDP-Q, as shown in Table 2(b). This contributed to reduce the total processing time by 14.6%. As a countermeasure to the above-mentioned "C2", LyDP-Q shared the look-up table of the M_1-level quantization with the M_2-level quantization instead of generating look-up tables for each quantization. This is because the look-up table of the M_1-level quantization includes that of the M_2-level quantization. As a result, LyDP-Q reduced the processing time of this process by 48.4% over SmDP-Q, as shown in Table 2(c). This contributed to reduce the total processing time by 6.2%.

Next, we analyze the complexity of the above-mentioned "C1" and "C2" for the approach based on DP quantization which is abbreviated as DP-Q. The complexity of "C1" is related to the number of feasible paths within the search range that is shared among both quantizations. In the following, feasible paths within the search range are abbreviated as candidate paths, and candidate paths that are shared between both quantizations are abbreviated as shared candidate paths. According to [17], the number of candidate paths for DP-Q that quantizes a K-level signal into a M-level signal is derived as follows:

$$\Omega_{\text{path}}(M,K) = \frac{1}{2}M^3 - \frac{2K+5}{2}M^2 + \frac{K^2+7K+4}{2}M - K^2 - K \quad (11)$$

The number of shared candidate paths for DP-Qs that output M_1-level/M_2-level signals from K-level signal is derived as follows:

$$\Omega_{\text{shared-path}}(M_1, M_2, K)$$

Table 2 Average processing time of SmDP-Q and LyDP-Q for all images (cells in the "reduction ratio" columns present the values yielded by Eq. (10), and cells in the SD rows present standard deviations of all channels).

(a) Total processing time

Channel	SmDP-Q [msec]	LyDP-Q [msec]	reduction ratio [%]
R	9371	7434	20.7
G	9370	7428	20.7
B	9376	7411	21.0
Ave.	9372	7424	20.8
SD	13	58	0.7

(b) Processing time for optimal path search

Channel	SmDP-Q [msec]	LyDP-Q [msec]	reduction ratio [%]
R	8169	6814	16.6
G	8167	6808	16.6
B	8170	6789	16.9
Ave.	8168	6803	16.7
SD	13	62	0.7

(c) Processing time for generating look-up table

Channel	SmDP-Q [msec]	LyDP-Q [msec]	reduction ratio [%]
R	1202	620	48.4
G	1204	620	48.5
B	1206	622	48.4
Ave.	1204	621	48.4
SD	10	5	0.1

$$= \left\{1 + \frac{1}{2}(K - M_2 + 2)(M_1 - 2)\right\}(K - M_2 + 1) \quad (12)$$

The derivation of the above equation can be found in Appendix.

Based on these analysis models, we found that complexity reduction obtained by eliminating "C1" is evaluated as $\Omega_{\text{shared-path}}(M_1, M_2, K)$; it reflects the complexity-related computations reused in searching for optimal paths of M_2-level quantization. Furthermore, the reduction ratio which is the above complexity reduction over the complexity relevant to this process in SmDP-Q is evaluated as follows:

$$CR_{\text{pash}}(M_1, M_2, K) = \frac{\Omega_{\text{shared-path}}(M_1, M_2, K)}{\Omega_{\text{path}}(M_1, K) + \Omega_{\text{path}}(M_2, K)} \quad (13)$$

For the case of $K = 4096$, $M_2 = 1024$ and $M_1 = 256$, the reduction ratio is 17.9%, which is close to the simulation results shown in the rightmost column in the "Ave." row of Table 2(b). Additionally, the standard deviations of the processing time were small as shown in the "SD" row in the table.

The complexity of "C2" is related to the number of candidate intervals in the input histogram, because the quantization error of each candidate interval needs to be computed and stored in $E[]$. According to [17], the number of candidate intervals for DP-Q that quantizes a K-level signal into a M-level signal is derived as follows:

$$\Omega_{\text{interval}}(M, K) = (-M^2 + M + K^2 + K)/2 \quad (14)$$

Based on this analysis model, we found that complexity reduction obtained by eliminating "C2" is evaluated as $\Omega_{\text{interval}}(M_2, K)$; the complexity for generating the look-up table of M_2-level quantization. Furthermore, the reduction ratio which is the above complexity reduction over the complexity relevant to this process in SmDP-Q is evaluated as follows:

$$CR_{\text{interval}}(M_1, M_2, K) = \frac{\Omega_{\text{interval}}(M_2, K)}{\Omega_{\text{interval}}(M_1, K) + \Omega_{\text{interval}}(M_2, K)} \quad (15)$$

For the case of $K = 4096$, $M_2 = 1024$ and $M_1 = 256$, the reduction ratio is 48.5%, which agrees well with the simulation results shown in the rightmost column in the "Ave." row of Table 2(c). Additionally, the standard deviations of the processing time were small as shown in the "SD" row in the table.

The above analysis for two-layer quantization can be generalized to two or more layers. Let L and M_l denote the number of layers and the quantization level of the l-th layer ($l = 1, \cdots, L$), respectively. Note that $M_1 < M_2 < \cdots < M_L$. In the case of an L-layer quantization, letting $\boldsymbol{M}^{(L)} = (M_1, \cdots, M_L)$, the reduction ratio of Eq. (13) is generalized to

$$CR_{\text{path}}(\boldsymbol{M}^{(L)}, K) = \frac{\sum_{l'=2}^{L} \Omega_{\text{shared-path}}(M_{l'-1}, M_{l'}, K)}{\sum_{l=1}^{L} \Omega_{\text{path}}(M_l, K)}$$

The above is derived in the followings. The number of candidate paths of LyDP-Q is $\Omega_{\text{path}}(M_1, K) + \sum_{l'=2}^{L} \left\{\Omega_{\text{path}}(M_{l'}, K) - \Omega_{\text{shared-path}}(M_{l'-1}, M_{l'}, K)\right\}$. On the other hand, the number of candidate paths of SmDP-Q is $\sum_{l=1}^{L} \Omega_{\text{path}}(M_l, K)$. So, the reduction of the former with respect to the latter is $\sum_{l'=2}^{L} \Omega_{\text{shared-path}}(M_{l'-1}, M_{l'}, K)$, and we obtain the above reduction ratio. Similarly, the reduction ratio of Eq. (15) is generalized to

$$CR_{\text{interval}}(\boldsymbol{M}^{(L)}, K) = \frac{\sum_{l'=2}^{L} \Omega_{\text{interval}}(M_{l'}, K)}{\sum_{l=1}^{L} \Omega_{\text{interval}}(M_l, K)}$$

This is because the candidate intervals of M_1 contain all candidate intervals of $M_2, \cdots, M_L$.

Table 3 provides comparisons between SmSDP-Q and LySDP-Q. Table 3(a) shows average processing time of SmSDP-Q and LySDP-Q for all images. The figures in the column "reduction ratio" in Table 3(a) were computed by applying the same concept as Eq. (10) to the candidate paths of SmSDP-Q and LySDP-Q. Table 3(b) and (c) give the processing time for searching for the optimal paths and that for generating look-up tables, respectively. As a result, LySDP-Q reduced the time taken by this process by 14.0% over SmSDP-Q, as shown in Table 3(b). This contributed to reducing the total processing time by 12.6%. LyDP-Q reduced the processing time of this process by 33.4% over

Table 3 Average processing time of SmSDP-Q and LySDP-Q for all images (cells in the "reduction ratio" column present values given by Eq. (10)).

(a) Total processing time

Channel	SmSDP-Q [msec]	LySDP-Q [msec]	reduction ratio [%]
R	1252	1054	15.9
G	1263	1061	16.0
B	1311	1097	16.3
Ave.	1276	1071	16.1

(b) Processing time for optimal path search

Channel	SmSDP-Q [msec]	LySDP-Q [msec]	reduction ratio [%]
R	1122	966	13.9
G	1131	973	14.0
B	1173	1006	14.3
Ave.	1142	982	14.0

(c) Processing time for generating look-up table

Channel	SmSDP-Q [msec]	LySDP-Q [msec]	reduction ratio [%]
R	130	87	32.9
G	132	88	33.1
B	138	91	34.0
Ave.	133	89	33.4

SmDP-Q, as shown in Table 3(c). This contributed to reducing the total processing time by 3.5%.

It was observed that the reduction ratio of LySDP-Q was less than that of LyDP-Q from the perspectives of total processing time, processing time of optimal path search, and processing time to generate look-up table.

We consider the cause of this observation by analyzing the complexity of "C1" and "C2" for approaches based on SDP-Q. Since SDP-Q skips insignificant elements in the histogram of the input signal, the number of insignificant elements impacts the complexity of approaches based on SDP-Q, such as SmSDP-Q and LySDP-Q. The number of significant elements varies depending on channels of images used in our experiments as shown in Fig. 7. This is why the processing times of SmSDP-Q and LySDP-Q depend on images in Fig. 6. In the following, let $\tilde{K}$ be the number of the significant elements. Based on the analysis models of Eqs. (11) and (12), the complexity reduction gained by eliminating "C1" is approximately evaluated as $\Omega_{\text{shared-path}}(M_1, M_2, \tilde{K})$. The reduction ratio which is the above complexity reduction over the complexity relevant to this process in SmSDP-Q is approximately evaluated as follows:

$$\frac{\Omega_{\text{shared-path}}(M_1, M_2, \tilde{K})}{\Omega_{\text{path}}(M_1, \tilde{K}) + \Omega_{\text{path}}(M_2, \tilde{K})} \tag{16}$$

The number of significant elements depends on the input signal. In the following, let us consider the complexity of SmSDP-Q and LySDP-Q using the average number of significant elements over all input signals (26 kinds of images) used in the experiments. The average number of the significant elements was 1597. In the case of $\tilde{K} = 1597$, $M_2 = 1024$ and $M_1 = 256$, the above analytical evaluation says that the reduction ratio would be 10.5%. This value is less than the reduction ratio of $K = 4096$, $M_2 = 1024$ and $M_1 = 256$, which was 48.5% as stated previously. This shows that the reduction ratio tends to decrease as the number of significant elements decreases. This trend closely resembles the experimental results. There was, however, some gap between the analytical evaluation and the simulation results shown in the rightmost column in Table 3(c). This is because Eqs. (12) and (11) are analytical models for DP-Q based approaches and can not exactly represent the number of candidate paths for SDP-Q. Specifically, the analytical models do not consider the complexity reduction achieved by restricting the search range, see Sect. 5.2 in [17].

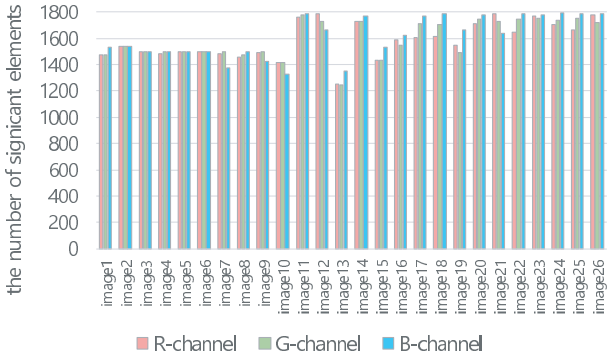


Fig. 7 The number of insignificant elements in each channel.

Based on this analysis model of Eq. (14), the complexity reduction achieved by eliminating "C2" is approximately $\Omega_{\text{interval}}(M_2, \tilde{K})$. The reduction ratio which is the above complexity reduction over the complexity relevant to this process in SmDP-Q is evaluated as follows:

$$\frac{\Omega_{\text{interval}}(M_2, \tilde{K})}{\Omega_{\text{interval}}(M_1, \tilde{K}) + \Omega_{\text{interval}}(M_2, \tilde{K})} \tag{17}$$

In the case of $\tilde{K} = 1597$, $M_2 = 1024$ and $M_1 = 256$, the above analytical evaluation says that the reduction ratio is 37.7%, which is close to simulation results shown in the rightmost column of Table 3(c).

4. Conclusion

This paper extended DP quantization so as to significantly reduce the complexity of multi-level quantization. The algorithm proposed herein eliminates the redundant computations between different quantization levels without losing the optimality of multi-level quantization. Experiments showed that the proposed algorithm attained 20.8% and 16.1% lower complexity (on average) than simultaneous quantization built on DP quantization and sparse DP quantization, respectively.

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, G. Ward, S. Pattanaik, P. Debevec, W. Heidrich, and K. Myszkowski, High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting, 2nd ed., Morgan Kaufmann Publisher, 2010.
[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," ISO/IEC JTC1/SC29/WG11/N16508, Oct. 2016.
[4] A. Artusi, T. Richter, T. Ebrahimi, and R.K. Mantiuk, "High dynamic range imaging technology [lecture notes]," IEEE Signal Process. Mag., vol.34, no.5, pp.165–172, Sept. 2017.
[5] 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 Technol., vol.26, no.1, pp.20–34, 2016.
[6] ISO/IEC, Information technology: Scalable compression and coding of continuous-tone still images – Part 1: Scalable compression and coding of continuous-tone still images, ISO/IEC 18477-1, 2015.
[7] ISO/IEC, Information technology: Scalable compression and coding of continuous-tone still images – Part 8: Lossless and nearlossless coding, ISO/IEC 18477-8, 2016.
[8] E. François, C. Fogg, Y. He, X. Li, A. Luthra, and A. Segall, "High dynamic range and wide color gamut video coding in HEVC: Status and potential future enhancements," IEEE Trans. Circuits Syst. Video Technol, vol.26, no.1, pp.63–75, 2016.
[9] ISO/IEC, Information technology: Scalable compression and coding of continuous-tone still images – Part 2: Coding of high dynamic range images, ISO/IEC 18477-2, 2016.
[10] ISO/IEC PDTR 23008-14: Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 14: Conversion and Coding Practices for HDR/WCG Y'CbCr 4:2:0 Video with PQ Transfer Characteristics, 2016.
[11] A.O. Zaid and A. Houimli, "HDR image compression with optimized JPEG coding," European Signal Processing Conference, pp.1539–1543, Aug. 2017.
[12] S.P. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inf. Theory, vol.IT-28, pp.129–136, March 1982.
[13] J. Max, "Quantizing for minimum distortion," IRE Trans. Inf. Theory, vol.IT-7, pp.7–12, March 1960.
[14] J.D. Bruce, Optimum quantizer, Ph.D. thesis, M.I.T., May 1964.
[15] 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.
[16] X. Wu, "Optimal quantization by matrix searching," J. Algorithms, vol.12, no.4, pp.663–673, Dec. 1991.
[17] Y. Bandoh, S. Takamura, and A. Shimizu, "Sparse DP quantization algorithm," IEICE Trans. Fundamentals, vol.E102-A, no.3, pp.553–565, March 2019.
[18] ITU-T and ISO/IEC JTC 1, High efficiency video coding, ITU-T Rec.H.265 and ISO/IEC 23008-2(HEVC), version 6, June 2019.
[19] ITU-T and ISO/IEC JTC 1, Advanced Video Coding for Generic Audio-Visual Services, ITU-T Rec.H.264 and ISO/IEC 14496-10 (AVC), version 13, June 2019.
[20] Y. Bandoh, S. Takamura, and A. Shimizu, "Complexity reduction of multi-level DP quantization through inter-level redundancy elimination," Proc. IEEE Int. Conf. Image Process., WP.PF.1, 2019.
[21] D. Muresan and M. Effros, "Quantization as histogram segmentation: Optimal scalar quantizer design in network systems," IEEE Trans. Inf. Theory, vol.54, no.1, pp.344–366, 2008.
[22] A.K. Khandani, "A hierarchical dynamic programming approach to fixed-rate, entropy-coded quantization," IEEE Trans. Inf. Theory, vol.42, no.4, pp.1298–1303, 1996.
[23] http://www.ite.or.jp/content/test-materials/uhdtv_a/
[24] http://www.ite.or.jp/content/test-materials/uhdtv_b/

Appendix: Derivation of Eq. (12)

Let us consider the candidate paths shared among two DPQs that output M_1-level/M_2-level signal from the same K-level signal. We evaluate the shared candidate paths in two cases:

(i) nodes in a range with $m = 0$
(ii) nodes in a range with $m = 1, \cdots, M_1 - 2$

Figure A·1 illustrates the case of $K = 7$, $M_2 = 5$ and $M_1 = 4$. The green nodes and the blue nodes represent the nodes in class (i) and class (ii), respectively. First, let us consider class (i). In this class, there are $K - M_2 + 1$ kinds of nodes, and every node has a single path. Accordingly, we find that there are $K - M_2 + 1$ kinds of paths. Figure A·1 shows the paths in this class as the green line segments. Second, let us consider class (ii). In this class, a node $(m + \ell, m)$ has $\ell + 1$ kinds of paths. In a group of nodes located at $(m + \ell, m)$ where $\ell = 0, 1, \cdots, K - M_2$, $m = 1, \cdots, M_1 - 2$, there are the following number of paths:

$$\sum_{m=1}^{M_1-2} \sum_{\ell=0}^{K-M_2} (\ell + 1) = (M_1 - 2) \sum_{\ell=0}^{K-M_2} (\ell + 1)$$

Figure A·1 shows the paths in this class as the blue line segments. From the above results, we obtain the following:

$$\begin{aligned}
&\Omega_{\text{shared-path}}(M, K) \\
&= (K - M_2 + 1) + (M_1 - 2) \sum_{\ell=0}^{K-M_2} (\ell + 1) \\
&= (K - M_2 + 1) + \frac{1}{2}(M_1 - 2)(K - M_2 + 2)(K - M_2 + 1) \\
&= \left\{1 + \frac{1}{2}(M_1 - 2)(K - M_2 + 2)\right\}(K - M_2 + 1)
\end{aligned}$$

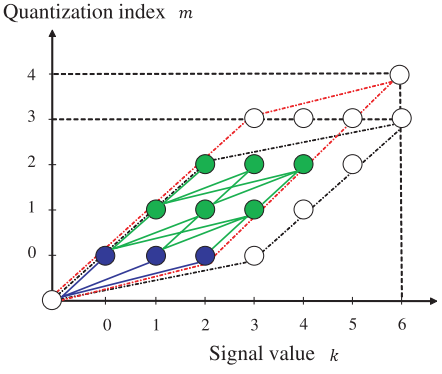


Fig. A·1 Shared nodes in the case of $K = 7$, $M_2 = 5$ and $M_1 = 4$.

Yukihiro Bandoh received the B.E., M.E., and Ph.D. degrees from Kyushu University, Japan, in 1996, 1998 and 2002, respectively. He granted JSPS Research Fellowship for Young Scientists from 2000 to 2002. In 2002, he joined Nippon Telegraph and Telephone (NTT) Corporation, where he has been engaged in research on efficient video coding for high realistic communication. He is currently a Distinguished Engineer at NTT Media intelligence Laboratories. Dr. Bandoh is a senior member of IEEE, IEICE, and IPSJ.

Seishi Takamura received the B.E., M.E. and Ph.D. degrees from the University of Tokyo in 1991, 1993 and 1996, respectively. In 1996 he joined Nippon Telegraph and Telephone (NTT) Corporation, where he has been engaged in research on efficient video coding and ultra-high quality video coding. He has fulfilled various duties in the research and academic community in current and prior roles including Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology, Executive Committee Member of IEEE Tokyo Section, Japan Council and Region 10, and Vice President-Industrial Relations and Development of APSIPA. He has also served as Chair of ISO/IEC JTC 1/SC 29 Japan National Body, Japan Head of Delegation of ISO/IEC JTC 1/SC 29, and as an International Steering Committee Member of the Picture Coding Symposium. From 2005 to 2006, he was a Visiting Scientist at Stanford University, California, USA. He is currently a Senior Distinguished Engineer at NTT Media Intelligence Laboratories. Dr. Takamura is a Fellow of IEEE, a senior member of IEICE and IPSJ, and a member of MENSA, APSIPA, SID and ITE.

Hideaki Kimata received the B.E. and M.E. degrees in applied physics, and the Ph.D. degree in electrical engineering respectively from Nagoya University, Nagoya, Japan, in 1993, 1995, and 2006. He joined Nippon Telegraph and Telephone Corporation (NTT) in 1995, and has been involved in the research and development of video coding, realistic communication, computer vision, and video recognition based on machine learning (deep learning). He is currently a Senior Research Engineer and also the Supervisor at NTT Media Intelligence Laboratories. He is a Chair of Technical Committee on Image Engineering of the Institute of Electronics, Information and Communication Engineers of Japan.