Available online at www.sciencedirect.com
SCIENCE DIRECT®
Signal Processing: *Image Communication* 19 (2004) 465–478
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Optimum word length allocation of integer DCT and its error analysis

Somchart Chokchaitam[a,*], Masahiro Iwahashi[b], Noriyoshi Kambayashi[b]

[a] *Department of Electrical Engineering, Thammasat University, Rangsit Campus, Pathum-Thani 12121, Thailand*
[b] *Department of Electrical Engineering, Nagaoka University of Technology, Nagaoka, Niigata 940-2188, Japan*

Received 25 September 2003; received in revised form 20 January 2004; accepted 22 March 2004

Abstract

Recently, the integer DCT (Int-DCT), which transforms an integer input to an integer output, is attracting many researchers' attention as an effective method for DCT-based lossy/lossless unified coding. So far, there are many reports relevant to the Int-DCT, but they have been limited to a few topics such as how to reduce the number of multipliers with the four-point lossless Hadamard transform and the non-separable two-dimensional Int-DCT. However, none of them is focused on how to express the multipliers' word length as short as possible for the reduction of hardware complexity.

In this report, we define a new "SNR sensitivity" as an indicator of how the word length truncation of multiplier coefficients affects quality of a reconstructed image. Based on the sensitivity, we propose a new word length allocation method. We also theoretically analyze errors in a reconstructed signal to confirm an effectiveness of the proposed method. As a result, the optimum word length allocation, which depends on a frequency spectrum of an input signal, significantly improves quality of a reconstructed image.

Keywords: Integer DCT; Optimum; Word Length Allocation; Error analysis

1. Introduction

The discrete cosine transform (DCT) is a well-known transform used in many international standards of image compression such as JPEG [6] and MPEG [4]. The DCT-based systems have huge advantage to image applications because they provide a high compression ratio. However, their coding systems are limited to operating in only lossy coding because distortion of decoded image is unavoidable with these lossy algorithms.

On the other hand, the integer transform [1], which includes rounding operations in the lifting structure [9], is becoming popular as a key technique to lossless and lossy unified waveform coding [11]. Especially the integer DCT [7,2,3] is attractive as the unified coding with compatibility to the conventional DCT-based algorithms. In Fig. 1, encoder applied the conventional lossy

*Corresponding author. Tel.: +66-9-4479300.
E-mail address: csomchar@engr.tu.ac.th (S. Chokchaitam).

0923-5965/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2004.03.001

DCT, whereas decoder applied the Int-DCT to illustrate its compatibility to the conventional DCT-based algorithms. Notice that the coding performance of the Int-DCT is similar to that of the conventional lossy DCT in a low bit-rate but it is slightly worse than that of the conventional lossy DCT in a high bit-rate because of rounding error discussed in Section 4.2.

So far, relevant to the integer DCT, previous reports focused on reducing the rounding operations with the non-separable 2D structuring [7] and reducing multipliers with the integer Hadamard transform [3]. Optimization of the basis function of the orthogonal transform (integer KLT) is also reported [10]. What seems to be lacking, however, is how to express multipliers' word length as short as possible for the reduction of hardware complexity.

In this report, we define a new "SNR sensitivity" as an indicator of how the word length truncation of multiplier coefficients affects quality of a reconstructed image. Based on the newly defined sensitivity, we propose a new word length allocation method. We also theoretically analyze errors in a reconstructed signal to confirm an effectiveness of the proposed method. This report is organized as follows. Overview of the integer DCT is summarized in Section 2. An error generated from finite word length allocation is theoretically analyzed in Section 3 and errors in a reconstructed signal are theoretically analyzed in Section 4. The "SNR sensitivity" is newly defined and applied to an optimum word length allocation using the least square method in Section 5. An effectiveness of the proposed method is confirmed in Section 6.

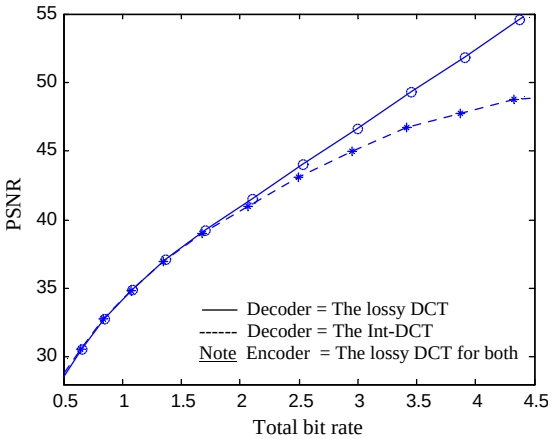

Fig. 1. Rate distortion curve of a decoded image "Barbara".

2. The integer DCT

2.1. The integer DCT (Int-DCT) [6–8]

Algorithm of the integer DCT (Int-DCT), illustrated in Fig. 2, is composed of the 4-point integer Hadamard transform (4-IHT) and integer rotation transform (IRT) described in Sections 2.2 and 2.3, respectively. The integer DCT transforms integer input vector $x(n)$, $(n = 0, 1, \ldots, 7)$ into integer output vector $y(n)$, $(n = 0, 1, \ldots, 7)$. Therefore, it is possible to achieve effective lossless coding by applying an entropy coding directly to the output vector. With inserting the quantization

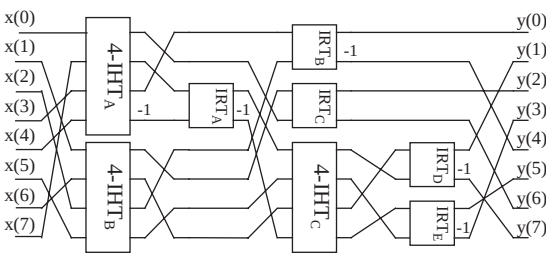

Fig. 2. The Integer DCT (forward transform).

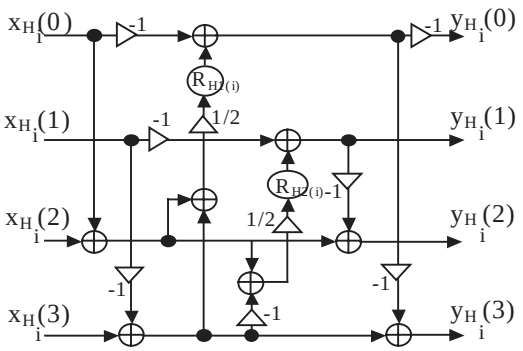

Fig. 3. The ith 4-point integer Hadamard transform (4-IHT$_i$) $(i = \mathrm{A}, \mathrm{B}, \mathrm{C})$.

procedure, its lossy coding is compatible to lossy coding of the conventional DCT-based algorithm such as JPEG [6] and MPEG [4].

2.2. The 4-point integer Hadamard transform (4-IHT)

The 4-point integer Hadamard transform (4-IHT) illustrated in Fig. 3, contains rounding operations "R" in the lifting structure. Rounding operation is a key part for an integer-to-integer transform. However, rounding operation generates an additive noise called "rounding effect" [8]. Relation between input and output of the ith 4-IHT ($i = \mathrm{A}, \mathrm{B}, \mathrm{C}$) is

$$\begin{bmatrix} y_{\mathrm{H}_i}(0) \\ y_{\mathrm{H}_i}(1) \\ y_{\mathrm{H}_i}(2) \\ y_{\mathrm{H}_i}(3) \end{bmatrix} = \mathbf{IHT}_i \begin{bmatrix} x_{\mathrm{H}_i}(0) \\ x_{\mathrm{H}_i}(1) \\ x_{\mathrm{H}_i}(2) \\ x_{\mathrm{H}_i}(3) \end{bmatrix} + \begin{bmatrix} -N_{\mathrm{H}_{1(i)}} \\ N_{\mathrm{H}_{2(i)}} \\ -N_{\mathrm{H}_{2(i)}} \\ -N_{\mathrm{H}_{1(i)}} \end{bmatrix}, \tag{1}$$

where $N_{\mathrm{H}_{1(i)}}$ and $N_{\mathrm{H}_{2(i)}}$ are additive noises generated from rounding effect in the ith 4-IHT [8] and matrix $\mathbf{IHT}_i$ is

$$\mathbf{IHT}_i = \begin{bmatrix} 0.5 & 0.5 & -0.5 & -0.5 \\ 0.5 & -0.5 & 0.5 & -0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & -0.5 & 0.5 \end{bmatrix}, \quad (\text{for } i = \mathrm{A}, \mathrm{B}, \mathrm{C}). \tag{2}$$

Similarly, relation between input and output of inverse of the ith 4-IHT ($i = \mathrm{A}, \mathrm{B}, \mathrm{C}$) is

$$\begin{bmatrix} x_{\mathrm{H}'_i}(0) \\ x_{\mathrm{H}'_i}(1) \\ x_{\mathrm{H}'_i}(2) \\ x_{\mathrm{H}'_i}(3) \end{bmatrix} = \mathbf{IHT}_i^{-1} \begin{bmatrix} y_{\mathrm{H}'_i}(0) \\ y_{\mathrm{H}'_i}(1) \\ y_{\mathrm{H}'_i}(2) \\ y_{\mathrm{H}'_i}(3) \end{bmatrix} + \begin{bmatrix} -N_{\mathrm{H}'_{1(i)}} \\ -N_{\mathrm{H}'_{2(i)}} \\ N_{\mathrm{H}'_{1(i)}} \\ -N_{\mathrm{H}'_{2(i)}} \end{bmatrix}, \tag{3}$$

where $x_{\mathrm{H}'_i}$ and $y_{\mathrm{H}'_i}$ are input and output of inverse of the ith 4-point integer Hadamard transform, respectively. $N_{\mathrm{H}'_{1(i)}}$ and $N_{\mathrm{H}'_{2(i)}}$ are additive noises generated from rounding effect in inverse of the ith 4-IHT and matrix $\mathbf{IHT}_i^{-1}$ is

$$\mathbf{IHT}_i^{-1} = \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & 0.5 & -0.5 \\ -0.5 & 0.5 & 0.5 & -0.5 \\ -0.5 & -0.5 & 0.5 & 0.5 \end{bmatrix} \quad (\text{for } i = \mathrm{A}, \mathrm{B}, \mathrm{C}). \tag{4}$$

2.3. Integer rotation transform (IRT)

As indicated in Fig. 2, the Int-DCT has five IRTs and each IRT has three multipliers. Therefore, the Int-DCT has 15 multipliers in total. Their values are given by

$$\mathbf{M}_{\mathrm{A}} = \mathbf{M}_{\mathrm{B}} = \begin{bmatrix} 1 - 2^{1/2} & 2^{-1/2} & 1 - 2^{1/2} \end{bmatrix},$$

$$\mathbf{M}_{\mathrm{C}} = \begin{bmatrix} \dfrac{\sin(\pi/8) - 1}{\cos(\pi/8)} & \cos(\pi/8) & \dfrac{\cos(3\pi/8) - 1}{\cos(\pi/8)} \end{bmatrix},$$

$$\mathbf{M}_{\mathrm{D}} = \begin{bmatrix} \dfrac{1 - \cos(3\pi/16)}{\sin(3\pi/16)} & -\sin(3\pi/16) & \dfrac{1 - \cos(3\pi/16)}{\sin(3\pi/16)} \end{bmatrix},$$

$$\mathbf{M}_{\mathrm{E}} = \begin{bmatrix} \dfrac{\cos(\pi/16) - 1}{\sin(\pi/16)} & \sin(\pi/16) & \dfrac{\cos(\pi/16) - 1}{\sin(\pi/16)} \end{bmatrix}, \tag{5}$$

where

$$\mathbf{M}_{(i)} = [m_{1(i)} \;\; m_{2(i)} \;\; m_{3(i)}] \; (i = \mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}).$$

Fig. 4 illustrates the integer rotation transform. Its input and output relation is

$$\begin{pmatrix} y_{\mathrm{R}_i}(0) \\ y_{\mathrm{R}_i}(1) \end{pmatrix} = \mathbf{IRT}_i \begin{pmatrix} x_{\mathrm{R}_i}(0) \\ x_{\mathrm{R}_i}(1) \end{pmatrix} + \begin{bmatrix} m_2 N_{m_{1(i)}} + N_{m_{2(i)}} \\ (1 + m_2 m_3) N_{m_{1(i)}} + m_3 N_{m_{2(i)}} + N_{m_{3(i)}} \end{bmatrix}, \tag{6}$$

where $m_{1(i)}$, $m_{2(i)}$, and $m_{3(i)}$ indicate multiplier coefficients in the ith IRT ($i = \mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}$). $N_{m_{1(i)}}$, $N_{m_{2(i)}}$ and $N_{m_{3(i)}}$ denote additive noise generated

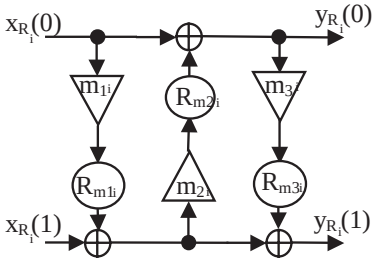

Fig. 4. The ith integer rotation transform (IRT_i) ($i = A, B, C, D, E$).

from rounding effect in ith IRT and matrix $\mathbf{IRT}_i$ is

$$\mathbf{IRT}_i = \begin{bmatrix} 1 + m_{1(i)}m_{2(i)} & m_{2(i)} \\ m_{1(i)} + m_{3(i)} + m_{1(i)}m_{2(i)}m_{3(i)} & 1 + m_{2(i)}m_{3(i)} \end{bmatrix}. \tag{7}$$

Similarly, relation between input and output of inverse of the integer rotation transform is

$$\begin{pmatrix} y_{R'_i}(0) \\ y_{R'_i}(1) \end{pmatrix} = \mathbf{IRT}_i^{-1} \begin{pmatrix} x_{R'_i}(0) \\ x_{R'_i}(1) \end{pmatrix} + \begin{bmatrix} N_{m'_{2(i)}} - m_2 N_{m'_{3(i)}} \\ N_{m'_{1(i)}} - m_1 N_{m'_{2(i)}} + (1 + m_1 m_2) N_{m'_{3(i)}} \end{bmatrix}, \tag{8}$$

where $x_{R'_i}$ and $y_{R'_i}$ are the input signal and output signal of inverse of the ith integer rotation transform, respectively. $N_{m'_{1(i)}}$, $N_{m'_{2(i)}}$ and $N_{m'_{3(i)}}$ are additive noise generated from rounding effect in inverse of ith IRT[] and matrix $\mathbf{IRT}_i^{-1}$ is

$$\mathbf{IRT}_i^{-1} = \begin{bmatrix} 1 + m_{2(i)}m_{3(i)} & -m_{2(i)} \\ -(m_{1(i)} + m_{3(i)} + m_{1(i)}m_{2(i)}m_{3(i)}) & 1 + m_{1(i)}m_{2(i)} \end{bmatrix}. \tag{9}$$

In an LSI implementation, each of these real numbers is approximated to a binary number with finite word length. Next, we analyze errors generated from finite word length allocation and errors in a reconstructed image in **3** and in **4**, respectively. The purpose of this report is how to allocate the optimum word length to these multiplier coefficients considering the "SNR sensitivity" in **5**.

3. An error generated from finite word length allocation

3.1. Finite word length expression

The multiplier coefficient $m_{j(i)}$, ($i = A, B, C, D, E$ and $j = 1, 2, 3$), is expressed as h_k, ($k = 0, 1, \ldots, 14$), by

$$h_k = (-1)^{B_0} \cdot \sum_{j=1}^{\infty} B_j 2^{-j}, \quad k = 0, 1, \ldots, 14, \tag{10}$$

where B_j ($j = 0, 1, \ldots$) is 0 or 1. Under the finite word length expression in this report, h_k is truncated into W_k [bit] binary value h'_k. Namely,

$$h'_k = (-1)^{B_0} \cdot \sum_{j=1}^{W_k} B'_j 2^{-j}, \quad k = 0, 1, \ldots, 14. \tag{11}$$

Value Δh_k is defined as a difference between value h_k and binary value h'_k as

$$\Delta h_k = h_k - h'_k. \tag{12}$$

3.2. An error generated from finite word length allocation

Considering errors generated from finite word length allocation, we can find an equivalent circuit of inverse of the integer rotation transform as illustrated in Fig. 5.

Errors generated from finite word length in inverse of the ith integer rotation transform are calculated from

$$\begin{bmatrix} N_{F0(i)} \\ N_{F1(i)} \end{bmatrix} = \mathbf{G}_{F_i} \begin{bmatrix} x_{R'_i}(0) \\ x_{R'_i}(1) \end{bmatrix}, \tag{13}$$

where

$$\begin{aligned} \mathbf{G}_{F_i} = {} & \mathbf{G}_{F_{1(i)}} \Delta h_{1(i)} + \mathbf{G}_{F_{2(i)}} \Delta h_{2(i)} \\ & + \mathbf{G}_{F_{3(i)}} \Delta h_{3(i)} + \mathbf{G}_{F_{4(i)}} \Delta h_{1(i)} \Delta h_{2(i)} \\ & + \mathbf{G}_{F_{5(i)}} \Delta h_{1(i)} \Delta h_{3(i)} \\ & + \mathbf{G}_{F_{6(i)}} \Delta h_{2(i)} \Delta h_{3(i)} \\ & + \mathbf{G}_{F_{7(i)}} \Delta h_{1(i)} \Delta h_{2(i)} \Delta h_{3(i)}. \end{aligned} \tag{14}$$

The $\Delta h_{1(i)}$, $\Delta h_{2(i)}$ and $\Delta h_{3(i)}$ parameters are small values, so the $\Delta h_{1(i)}\Delta h_{2(i)}$, $\Delta h_{1(i)}\Delta h_{3(i)}$, $\Delta h_{2(i)}\Delta h_{3(i)}$ and $\Delta h_{1(i)}\Delta h_{2(i)}\Delta h_{3(i)}$ parameters are close to zero.

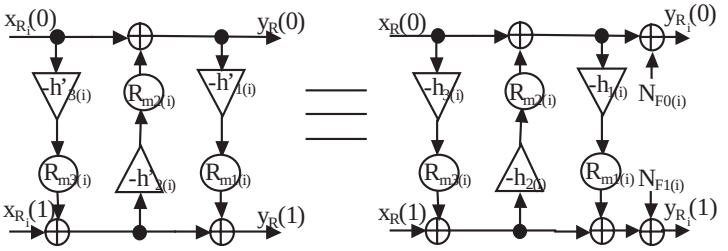

Fig. 5. An equivalent circuit of inverse of the ith integer rotation transform (IRT) ($i = \mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}$).

Therefore, parameter $\mathbf{G}_{\mathrm{F}}$ can be approximated as

$$\mathbf{G}_{\mathrm{F}_i} \approx \mathbf{G}_{\mathrm{F}_{1(i)}} \Delta h_{1(i)} + \mathbf{G}_{\mathrm{F}_{2(i)}} \Delta h_{2(i)} + \mathbf{G}_{\mathrm{F}_{3(i)}} \Delta h_{3(i)}, \quad (15)$$

where

$$\mathbf{G}_{\mathrm{F}_{1(i)}} = \begin{bmatrix} 0 & 0 \\ -(1 + h_{2(i)} h_{3(i)}) & h_{2(i)} \end{bmatrix},$$

$$\mathbf{G}_{\mathrm{F}_{2(i)}} = \begin{bmatrix} h_{3(i)} & -1 \\ -h_{1(i)} h_{3(i)} & h_{1(i)} \end{bmatrix},$$

$$\mathbf{G}_{\mathrm{F}_{3(i)}} = \begin{bmatrix} h_{2(i)} & 0 \\ -(1 + h_{1(i)} h_{2(i)}) & 0 \end{bmatrix}. \quad (16)$$

4. Analysis on errors in a reconstructed signal

In this section, we analyze errors between an original signal and a reconstructed signal. A variance of the errors (σ_E^2) is calculated from

$$\sigma_E^2 = \frac{1}{N} \sum_{n=0}^{N-1} \{x'(n) - x(n)\}^2, \quad (17)$$

where $x(n)$ and $x'(n)$ denote an original signal and a reconstructed signal, respectively. "n" denotes a sequence of input signal where "n"$= 0, 1, 2, \ldots, N-1$.

4.1. Analysis on the integer DCT

By neglecting additional noises in the previous section, we can write output ($\mathbf{Y}$) of the integer DCT in term of its input ($\mathbf{X}$) as

$$\mathbf{Y} = \mathbf{IDCT} \cdot \mathbf{X}$$

$$\mathbf{X} = [x(0)\ x(1)\ x(2)\ x(3)\ x(4)\ x(5)\ x(6)\ x(7)]^{\mathrm{T}},$$

$$\mathbf{Y} = [y(0)\ y(1)\ y(2)\ y(3)\ y(4)\ y(5)\ y(6)\ y(7)]^{\mathrm{T}}, \quad (18)$$

The **IDCT** matrix denotes a integer-DCT-transform matrix as

$$\mathbf{IDCT} = \mathbf{P}_5 \cdot \mathbf{S}_4 \cdot \mathbf{P}_4 \cdot \mathbf{S}_3 \cdot \mathbf{P}_3 \cdot \mathbf{S}_2 \cdot \mathbf{P}_2 \cdot \mathbf{S}_1 \cdot \mathbf{P}_1, \quad (19)$$

where P_i matrices ($i = 1, 2, 3, 4, 5$) and S_j matrices ($j = 1, 2, 3, 4$) denote permutation matrices and transform matrices of the Int-DCT, respectively. P_i matrices ($i = 1, 2, 3, 4, 5$) are

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

$$\mathbf{P}_2 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

$$\mathbf{P}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{P}_4=\begin{bmatrix}1&0&0&0&0&0&0&0\\0&-1&0&0&0&0&0&0\\0&0&1&0&0&0&0&0\\0&0&0&1&0&0&0&0\\0&0&0&0&0&0&1&0\\0&0&0&0&1&0&0&0\\0&0&0&0&0&0&0&1\\0&0&0&0&0&1&0&0\end{bmatrix},$$

$$\mathbf{P}_5=\begin{bmatrix}1&0&0&0&0&0&0&0\\0&0&0&0&1&0&0&0\\0&0&1&0&0&0&0&0\\0&0&0&0&0&0&0&-1\\0&1&0&0&0&0&0&0\\0&0&0&0&0&0&1&0\\0&0&0&1&0&0&0&0\\0&0&0&0&0&-1&0&0\end{bmatrix}, \tag{20}$$

and S_j matrices ($j=1,2,3,4$) are

$$\mathbf{S}_1=\begin{bmatrix}\mathbf{IHT}_A&\mathbf{Z}_4\\\mathbf{Z}_4&\mathbf{IHT}_B\end{bmatrix},\quad \mathbf{S}_2=\begin{bmatrix}\mathbf{T}_1&\mathbf{Z}_4\\\mathbf{Z}_4&\mathbf{I}_4\end{bmatrix},$$
$$\mathbf{S}_3=\begin{bmatrix}\mathbf{T_2}&\mathbf{Z}_4\\\mathbf{Z}_4&\mathbf{IHT}_C\end{bmatrix},\quad \mathbf{S}_4=\begin{bmatrix}\mathbf{I}_4&\mathbf{Z}_4\\\mathbf{Z}_4&\mathbf{T}_3\end{bmatrix}, \tag{21}$$

where

$$\mathbf{Z}_4=\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix},\quad \mathbf{I}_4=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix},$$

$$\mathbf{T}_1=\begin{bmatrix}\mathbf{I}_2&\mathbf{Z}_2\\\mathbf{Z}_2&\mathbf{IRT}_A\end{bmatrix},\quad \mathbf{I}_2=\begin{bmatrix}1&0\\0&1\end{bmatrix},\mathbf{Z}_2=\begin{bmatrix}0&0\\0&0\end{bmatrix},$$
$$\mathbf{T}_2=\begin{bmatrix}\mathbf{IRT}_B&\mathbf{Z}_2\\\mathbf{Z}_2&\mathbf{IRT}_C\end{bmatrix},\quad \mathbf{T}_3=\begin{bmatrix}\mathbf{IRT}_D&\mathbf{Z}_2\\\mathbf{Z}_2&\mathbf{IRT}_E\end{bmatrix}. \tag{22}$$

Similarly, we can write output of the integer DCT (x') in terms of its input (y') as

$$\mathbf{X}'=\mathbf{IDCT}^{-1}\cdot\mathbf{Y}',$$
$$\mathbf{X}'=[x'(0)\ x'(1)\ x'(2)\ x'(3)\ x'(4)\ x'(5)\ x'(6)\ x'(7)]^{\mathrm{T}},$$
$$\mathbf{Y}'=[y'(0)\ y'(1)\ y'(2)\ y'(3)\ y'(4)\ y'(5)\ y'(6)\ y'(7)]^{\mathrm{T}}. \tag{23}$$

The $\mathbf{IDCT}^{-1}$ matrix denotes an inverse of the Int-DCT transform matrix as

$$\mathbf{IDCT}^{-1}=\mathbf{P}_1^{-1}\cdot\mathbf{S}_1^{-1}\cdot\mathbf{P}_2^{-1}\cdot\mathbf{S}_2^{-1}\cdot\mathbf{P}_3^{-1}\cdot\mathbf{S}_3^{-1}\cdot\mathbf{P}_4^{-1}\cdot\mathbf{S}_4^{-1}\cdot\mathbf{P}_5^{-1}, \tag{24}$$

where

$$\mathbf{P}_i^{-1}=\mathbf{P}_i^{\mathrm{T}};\mathbf{S}_1^{-1}=\begin{bmatrix}\mathbf{IHT}_A^{-1}&\mathbf{Z}_4\\\mathbf{Z}_4&\mathbf{IHT}_B^{-1}\end{bmatrix},$$
$$\mathbf{S}_2^{-1}=\begin{bmatrix}\mathbf{T}_1^{-1}&\mathbf{Z}_4\\\mathbf{Z}_4&\mathbf{I}_4\end{bmatrix},\quad \mathbf{S}_3^{-1}=\begin{bmatrix}\mathbf{T}_2^{-1}&\mathbf{Z}_4\\\mathbf{Z}_4&\mathbf{IHT}_C^{-1}\end{bmatrix}, \tag{25}$$
$$\mathbf{S}_4^{-1}=\begin{bmatrix}\mathbf{I}_4&\mathbf{Z}_4\\\mathbf{Z}_4&\mathbf{T}_3^{-1}\end{bmatrix},\quad \mathbf{T}_1^{-1}=\begin{bmatrix}\mathbf{I}_2&\mathbf{Z}_2\\\mathbf{Z}_2&\mathbf{IRT}_A^{-1}\end{bmatrix},$$
$$\mathbf{T}_2^{-1}=\begin{bmatrix}\mathbf{IRT}_B^{-1}&\mathbf{Z}_2\\\mathbf{Z}_2&\mathbf{IRT}_C^{-1}\end{bmatrix},\quad \mathbf{T}_3^{-1}=\begin{bmatrix}\mathbf{IRT}_D^{-1}&\mathbf{Z}_2\\\mathbf{Z}_2&\mathbf{IRT}_E^{-1}\end{bmatrix}.$$

4.2. Errors in a reconstructed signal

In this report, we consider the coding system, where the Int-DCT is applied to both the encoder and decoder. If the same coefficients are used between encoder and decoder, there is no error on the decoded image caused by finite word length expression. However, we do not focus this case. Next, we analyze in case the same coefficients are not used between encoder and decoder.

Therefore, errors in a reconstructed signal in the following analysis are generated from three different sources: quantization, rounding operation and finite word length allocation. We can write the errors ($\Delta\mathbf{x}$) in a reconstructed signal as

$$\Delta\mathbf{x}=\mathbf{x}'-\mathbf{x}=\mathbf{N}_{\mathrm{TQ}}+\mathbf{N}_{\mathrm{TR}}+\mathbf{N}_{\mathrm{TF}}, \tag{26}$$

where $\mathbf{N}_{\mathrm{TQ}}$, $\mathbf{N}_{\mathrm{TR}}$, and $\mathbf{N}_{\mathrm{TF}}$ denote errors generated from quantization, rounding operation and finite word length allocation, respectively.

(A) *Quantization errors*:

Quantization errors ($\mathbf{N}_{\mathrm{TQ}}$) are calculated from

$$\mathbf{N}_{\mathrm{TQ}}=\mathbf{IDCT}^{-1}\cdot\mathbf{N}_{\mathbf{Q}} \tag{27}$$

$$\mathbf{N}_{\mathbf{Q}}=[N_{\mathrm{Q}_0}\ N_{\mathrm{Q}_1}\ N_{\mathrm{Q}_2}\ N_{\mathrm{Q}_3}\ N_{\mathrm{Q}_4}\ N_{\mathrm{Q}_5}\ N_{\mathrm{Q}_6}\ N_{\mathrm{Q}_7}]^{\mathrm{T}},$$

where N_{Qi} denotes quantization errors in subband ith. The errors generated from quantization ($\mathbf{N}_{\mathrm{TQ}}$) depend on a value of quantization step size [5].

For example, if quantization step size is big, variance of its error is big too.

(B) *Rounding errors*:

Next, we calculate rounding errors ($\mathbf{N}_{TR}$) from

$$\begin{aligned}\mathbf{N}_{TR} = {} & \mathbf{P}_1^{-1}\cdot\mathbf{N}_{R_1'} + \mathbf{SP}_2^{-1}\cdot\mathbf{N}_{R_2'} + \mathbf{SP}_3^{-1}\cdot\mathbf{N}_{R_3'} \\ & + \mathbf{SP}_4^{-1}\cdot\mathbf{N}_{R_4'} + \mathbf{SP}_4^{-1}\cdot\mathbf{N}_{R_4} \\ & + \mathbf{SP}_3^{-1}\cdot\mathbf{N}_{R_3} + \mathbf{SP}_2^{-1}\cdot\mathbf{N}_{R_2} \\ & + \mathbf{P}_1^{-1}\cdot\mathbf{N}_{R_1}, \end{aligned} \tag{28}$$

where $\mathbf{SP}_i^{-1}$ $(i = 2,3,4)$ denotes a transform matrix calculated from

$$\mathbf{SP}_2^{-1} = \mathbf{P}_1^{-1}\cdot\mathbf{S}_1^{-1}\cdot\mathbf{P}_2^{-1},$$

$$\mathbf{SP}_3^{-1} = \mathbf{P}_1^{-1}\cdot\mathbf{S}_1^{-1}\cdot\mathbf{P}_2^{-1}\cdot\mathbf{S}_2^{-1}\cdot\mathbf{P}_3^{-1},$$

$$\mathbf{SP}_4^{-1} = \mathbf{P}_1^{-1}\cdot\mathbf{S}_1^{-1}\cdot\mathbf{P}_2^{-1}\cdot\mathbf{S}_2^{-1}\cdot\mathbf{P}_3^{-1}\cdot\mathbf{S}_3^{-1}\cdot\mathbf{P}_4^{-1} \tag{29}$$

and $\mathbf{N}_{R_i}$ and $\mathbf{N}_{R_i'}$ denote errors generated from rounding operations in forward transform and inverse transform of the Int-DCT, respectively. The $\mathbf{N}_{R_i}$ and $\mathbf{N}_{R_i'}$ are

$$\begin{aligned}\mathbf{N}_{R_1'} = {} & [-N_{H_{1A}'} \; -N_{H_{2A}'} \; N_{H_{1A}'} \; -N_{H_{2A}'} \; -N_{H_{1B}'} \\ & -N_{H_{2B}'} \; N_{H_{1B}'} \; -N_{H_{2B}'}]^T, \end{aligned}$$

$$\mathbf{N}_{R_2'} = [0 \; 0 \; N_{IR_{1A}} \; N_{IR_{2A}} \; 0 \; 0 \; 0 \; 0]^T,$$

$$\begin{aligned}\mathbf{N}_{R_3'} = {} & [N_{IR_{1B}} \; N_{IR_{2B}} \; N_{IR_{1C}} \; N_{IR_{2C}} \; -N_{H_{1C}'} \\ & -N_{H_{2C}'} \; N_{H_{1C}'} \; -N_{H_{2C}'}]^T, \end{aligned} \tag{30}$$

$$\mathbf{N}_{R_4'} = [0 \; 0 \; 0 \; 0 \; N_{IR_{1D}} \; N_{IR_{2D}} \; N_{IR_{1E}} \; N_{IR_{2E}}]^T,$$

$$\mathbf{N}_{R_4} = [0 \; 0 \; 0 \; 0 \; N_{FR_{1D}} \; N_{FR_{2D}} \; N_{FR_{1E}} \; N_{FR_{2E}}]^T,$$

$$\begin{aligned}\mathbf{N}_{R_3} = {} & [N_{FR_{1B}} \; N_{FR_{2B}} \; N_{FR_{1C}} \; N_{FR_{2C}} \\ & -N_{H_{1C}} \; N_{H_{2C}} \; -N_{H_{1C}} \; -N_{H_{2C}}]^T, \end{aligned}$$

$$\mathbf{N}_{R_2} = [0 \; 0 \; N_{FR_{1A}} \; N_{FR_{2A}} \; 0 \; 0 \; 0 \; 0]^T,$$

$$\begin{aligned}\mathbf{N}_{R_1} = {} & [-N_{H_{1A}} \; N_{H_{2A}} \; -N_{H_{1A}} \; -N_{H_{2A}} \\ & -N_{H_{1B}} \; N_{H_{2B}} \; -N_{H_{1B}} \; -N_{H_{2B}}]^T, \end{aligned}$$

where

$$N_{IR_{1(i)}} = -m_2 N_{m_{3(i)}'} + N_{m_{2(i)}'},$$

$$N_{IR_{2(i)}} = (1 + m_1 m_2) N_{m_{3(i)}'} - m_1 N_{m_{2(i)}'} + N_{m_{1(i)}'},$$

$$N_{FR_{1(i)}} = m_2 N_{m_{1(i)}} + N_{m_{2(i)}}, \tag{31}$$

$$N_{FR_{2(i)}} = (1 + m_2 m_3) N_{m_{1(i)}} + m_3 N_{m_{2(i)}} + N_{m_{3(i)}}.$$

The N_H, $N_{H'}$, N_m and $N_{m'}$ denote additive noises in Eqs. (1), (3), (6) and (8), respectively. Errors generated from rounding operation ($\mathbf{N}_{TR}$) do not directly depend on a value of quantization step size, so their variance is quite constant. However, if a value of quantization step size equals one (or quantization part is not applied), rounding errors from inverse transform will compensate with rounding errors from forward transform. Moreover, a variance of rounding error is very small comparing to variances of the other errors.

(C) *Finite-word-length errors*: Finally, we calculate errors generated from finite word length allocation ($\mathbf{N}_{TF}$) from

$$\begin{aligned}\mathbf{N}_{TF} = {} & (\mathbf{F}_1^{-1}\cdot\mathbf{N}_{F1}\cdot\mathbf{F}_1 + \mathbf{F}_2^{-1}\cdot\mathbf{N}_{F2}\cdot\mathbf{F}_2 \\ & + \mathbf{F}_3^{-1}\cdot\mathbf{N}_{F3}\cdot\mathbf{F}_3)\cdot\mathbf{X}, \end{aligned} \tag{32}$$

where

$$\mathbf{F}_1^{-1} = \mathbf{P}_1^{-1}\cdot\mathbf{S}_1^{-1}\cdot\mathbf{P}_2^{-1},$$

$$\mathbf{F}_2^{-1} = \mathbf{P}_1^{-1}\cdot\mathbf{S}_1^{-1}\cdot\mathbf{P}_2^{-1}\cdot\mathbf{S}_2^{-1}\cdot\mathbf{P}_3^{-1},$$

$$\mathbf{F}_3^{-1} = \mathbf{P}_1^{-1}\cdot\mathbf{S}_1^{-1}\cdot\mathbf{P}_2^{-1}\cdot\mathbf{S}_2^{-1}\cdot\mathbf{P}_3^{-1}\cdot\mathbf{S}_3^{-1}\cdot\mathbf{P}_4^{-1},$$

$$\mathbf{F}_1 = \mathbf{S}_2\cdot\mathbf{P}_2\cdot\mathbf{S}_1\cdot\mathbf{P}_1,$$

$$\mathbf{F}_3 = \mathbf{S}_3\cdot\mathbf{P}_3\cdot\mathbf{S}_2\cdot\mathbf{P}_2\cdot\mathbf{S}_1\cdot\mathbf{P}_1,$$

$$\mathbf{F}_3 = \mathbf{S}_4\cdot\mathbf{P}_4\cdot\mathbf{S}_3\cdot\mathbf{P}_3\cdot\mathbf{S}_2\cdot\mathbf{P}_2\cdot\mathbf{S}_1\cdot\mathbf{P}_1, \tag{33}$$

$$\mathbf{N}_{F1} = \begin{bmatrix}\mathbf{GF}_1 & \mathbf{Z}_4 \\ \mathbf{Z}_4 & \mathbf{Z}_4\end{bmatrix}; \quad \mathbf{GF}_1 = \begin{bmatrix}\mathbf{Z}_2 & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{F_A}\end{bmatrix},$$

$$\mathbf{N}_{F2} = \begin{bmatrix}\mathbf{GF}_2 & \mathbf{Z}_4 \\ \mathbf{Z}_4 & \mathbf{Z}_4\end{bmatrix}; \quad \mathbf{GF}_2 = \begin{bmatrix}\mathbf{G}_{F_B} & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{F_C}\end{bmatrix},$$

$$\mathbf{N}_{F3} = \begin{bmatrix}\mathbf{Z}_4 & \mathbf{Z}_4 \\ \mathbf{Z}_4 & \mathbf{GF}_3\end{bmatrix}; \quad \mathbf{GF}_3 = \begin{bmatrix}\mathbf{G}_{F_D} & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{F_E}\end{bmatrix},$$

where $\mathbf{G}_{F_i}$ $(i = A, B, C, D, E)$ denote parameter $\mathbf{G}_{F_i}$ in Eq. (15). The errors generated from finite word

length allocation (N_{TF}) depend on number of W_k [bit] in Eq. (11).

4.3. An approximated variance of the errors

From Eq. (26), we illustrated errors in a reconstructed signal written in terms of errors generated from three different sources: quantization, rounding operation and finite word length allocation. In this section, we theoretically approximate a variance of the errors based on three coding conditions: no quantization, a big quantization step size and a small quantization step size.

(A) *No quantization*: In this condition, quantization is not applied, so there is no quantization error. Moreover, there is no rounding error because rounding errors in inverse transform cancel to rounding errors in forward transform.

Therefore, we can approximate a variance of the errors for this condition to be

$$\boldsymbol{\sigma}_{\mathrm{E}}^2 = \boldsymbol{\sigma}_{\mathrm{N_{TF}}}^2. \tag{34}$$

(B) *A big quantization step size*: In this case, the errors are generated from three sources. However, variances of quantization errors and finite-word-length error are very big compared to a variance of rounding error. If we assume that errors generated from each source are independent, we can calculate a variance of the errors in this case from

$$\boldsymbol{\sigma}_{\mathrm{E}}^2 \approx \boldsymbol{\sigma}_{\mathrm{N_{TQ}}}^2 + \boldsymbol{\sigma}_{\mathrm{N_{TF}}}^2. \tag{35}$$

(C) *A small quantization step size*: In this case, the error generated from finite-word-length dominates the other errors. Therefore, we can approximate a variance of the error signal in this case to be

$$\boldsymbol{\sigma}_{\mathrm{E}}^2 \approx \boldsymbol{\sigma}_{\mathrm{N_{TF}}}^2. \tag{36}$$

5. The proposed optimum word length allocation

5.1. The SNR sensitivity

From Eqs. (15), (16), (32), and (33), we can rewrite errors generated from finite word length allocation ($\mathbf{N}_{\mathrm{TF}}$) as

$$\mathbf{N}_{\mathrm{TF}} = \sum_{k=0}^{14} (\mathbf{S}_{Hk} \cdot \Delta h_k), \tag{37}$$

where the $\mathbf{S}_{Hk}$ called "SNR sensitivity" is defined as an effect of the finite word length expression on a quality of the decoded image.

$$\mathbf{S}_{\mathrm{H}(j)} = \begin{cases} \mathbf{F}_1^{-1} \cdot \mathbf{N}_{\mathrm{HF}(j)} \cdot \mathbf{F}_1 \cdot \mathbf{X} & \text{for } j = 0,1,2 \\ \mathbf{F}_2^{-1} \cdot \mathbf{N}_{\mathrm{HF}(j)} \cdot \mathbf{F}_2 \cdot \mathbf{X} & \text{for } j = 3,4,5,6,7,8 \\ \mathbf{F}_3^{-1} \cdot \mathbf{N}_{\mathrm{HF}(j)} \cdot \mathbf{F}_3 \cdot \mathbf{X} & \text{for } j = 9,10,11,12,13,14, \end{cases} \tag{38}$$

where

$$\mathbf{N}_{\mathrm{HF}(j)} = \begin{cases} \begin{bmatrix} \mathbf{HF}_j & \mathbf{Z}_4 \\ \mathbf{Z}_4 & \mathbf{Z}_4 \end{bmatrix} & \text{for } j = 0,1,2,\ldots,8, \\ \begin{bmatrix} \mathbf{Z}_4 & \mathbf{Z}_4 \\ \mathbf{Z}_4 & \mathbf{HF}_j \end{bmatrix} & \text{for } j = 9,10,11,\ldots,14, \end{cases} \tag{39}$$

$$\begin{aligned} &\mathbf{HF}_0 = \begin{bmatrix} \mathbf{Z}_2 & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{\mathrm{F_{1A}}} \end{bmatrix}, \quad \mathbf{HF}_1 = \begin{bmatrix} \mathbf{Z}_2 & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{\mathrm{F_{2A}}} \end{bmatrix}, \\ &\mathbf{HF}_2 = \begin{bmatrix} \mathbf{Z}_2 & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{\mathrm{F_{3A}}} \end{bmatrix}, \mathbf{HF}_3 = \begin{bmatrix} \mathbf{G}_{\mathrm{F_{1B}}} & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{Z}_2 \end{bmatrix}, \\ &\mathbf{HF}_4 = \begin{bmatrix} \mathbf{G}_{\mathrm{F_{1B}}} & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{Z}_2 \end{bmatrix}, \\ &\mathbf{HF}_5 = \begin{bmatrix} \mathbf{G}_{\mathrm{F_{1B}}} & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{Z}_2 \end{bmatrix}, \quad \mathbf{HF}_6 = \begin{bmatrix} \mathbf{Z}_2 & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{\mathrm{F_{1C}}} \end{bmatrix}, \\ &\mathbf{HF}_7 = \begin{bmatrix} \mathbf{Z}_2 & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{\mathrm{F_{2C}}} \end{bmatrix}, \quad \mathbf{HF}_8 = \begin{bmatrix} \mathbf{Z}_2 & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{\mathrm{F_{3C}}} \end{bmatrix}, \\ &\mathbf{HF}_9 = \begin{bmatrix} \mathbf{G}_{\mathrm{F_{1D}}} & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{Z}_2 \end{bmatrix}, \\ &\mathbf{HF}_{10} = \begin{bmatrix} \mathbf{G}_{\mathrm{F_{1D}}} & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{Z}_2 \end{bmatrix}, \quad \mathbf{HF}_{11} = \begin{bmatrix} \mathbf{G}_{\mathrm{F_{1D}}} & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{Z}_2 \end{bmatrix}, \\ &\mathbf{HF}_{12} = \begin{bmatrix} \mathbf{Z}_2 & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{\mathrm{F_{1E}}} \end{bmatrix}, \quad \mathbf{HF}_{13} = \begin{bmatrix} \mathbf{Z}_2 & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{\mathrm{F_{2E}}} \end{bmatrix}, \\ &\mathbf{HF}_{14} = \begin{bmatrix} \mathbf{Z}_2 & \mathbf{Z}_2 \\ \mathbf{Z}_2 & \mathbf{G}_{\mathrm{F_{3E}}} \end{bmatrix}. \end{aligned} \tag{40}$$

The $\mathbf{G}_{\mathrm{F}_{1(i)}}$ $(i = \mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E})$ denotes $\mathbf{G}_{\mathrm{F}_{1(i)}}$ parameter written in Eq. (15). From Eq. (37), we can calculate a variance of finite-word-length

errors from

$$\sigma^2_{N_{TF}} = \sum_{k=0}^{14} \{||\mathbf{S}_{Hk}||^2 \cdot (\Delta h_k)^2\}, \quad (41)$$

where $||\mathbf{S}_{Hk}||^2$ is calculated from

$$||\mathbf{S}_{Hk}||^2 = \frac{1}{8} \sum_{j=0}^{7} S^2_{Hk}[j]. \quad (42)$$

We also define the "relative SNR sensitivity" by

$$SR_k = \frac{||\mathbf{S}_{Hk}||}{\prod_{p=0}^{14} \sqrt[15]{||\mathbf{S}_{Hp}||}}, \quad k = 0, 1, \ldots, 14. \quad (43)$$

Next, we will illustrate a relation between the SNR sensitivity and PSNR defined as

$$PSNR = 10 \log_{10}\left(\frac{255^2}{\sigma_E^2}\right)[dB]. \quad (44)$$

If quantization is not applied, errors are generated from only finite-word-length errors. From Eqs. (34), (41) and (43), we can rewrite PSNR as

$$\begin{aligned} PSNR &= 10 \log_{10}\left(\frac{255^2}{\sum_{k=0}^{14}\{||\mathbf{S}_{Hk}||^2 \cdot (\Delta h_k)^2\}}\right) \\ &= 20 \log_{10} 255 \\ &\quad - 10 \log_{10}\left(\sum_{k=0}^{14}\{||\mathbf{S}_{Hk}||^2 \cdot (\Delta h_k)^2\}\right). \end{aligned} \quad (45)$$

If the finite word length allocation is applied to only h_j, so Eq. (45) can be reduced to

$$\begin{aligned} PSNR &= 20 \log_{10} 255 - 20 \log_{10}||\mathbf{S}_{Hj}|| \\ &\quad - 20 \log_{10}(\Delta h_j). \end{aligned} \quad (46)$$

We can calculate a value of $||\mathbf{S}_{Hj}||$, if we know a frequency spectrum of an input signal. From Eq. (46), we can simply write a relation between PSNR and $\log_{10}(\Delta h_j)$ as

$$PSNR = c_{0(j)} + c_{1(j)} \cdot \log_{10}(\Delta h_j) \quad (47)$$

$$c_{0(j)} = 20 \log_{10} 255 - 20 \log_{10}||\mathbf{S}_{Hj}||, \quad c_{1(j)} = -20.$$

On the other hand, a relation between the sensitivity $||\mathbf{S}_{Hj}||$ and $c_{0(j)}$ is

$$||\mathbf{S}_{Hj}|| = 255 \cdot 10^{-c_{0(j)/20}}. \quad (48)$$

In Table 1, we theoretically calculate the SNR sensitivity $||\mathbf{S}_{Hk}||$ by applying the AR(1) model with correlation coefficient ρ as an input signal,

Table 1
The theoretical SNR sensitivity $||\mathbf{S}_{Hk}||$ based on the AR(1) model

k	i	j	$\|\mathbf{S}_{Hk}\|$			
			$\rho = 0.95$	$\rho = 0.9$	$\rho = 0.85$	$\rho = 0.8$
0		1	9.09	18.46	28.43	39.45
1	A	2	11.47	23.35	36.17	50.59
2		3	1.60	3.20	4.80	6.40
3		1	88.95	88.88	90.52	94.19
4	B	2	56.14	55.30	54.87	54.87
5		3	125.62	125.00	126.36	130.04
6		1	0.23	0.96	2.25	4.17
7	C	2	0.65	2.71	6.37	11.96
8		3	0.73	3.04	7.14	13.36
9		1	12.10	24.49	37.52	51.67
10	D	2	4.76	9.65	14.83	20.51
11		3	14.63	29.62	45.41	62.57
12		1	0.76	1.61	2.68	4.13
13	E	2	1.51	3.21	5.32	8.14
14		3	0.46	0.99	1.65	2.55

with its frequency spectrum is

$$|X(e^{j\omega})| = \frac{1-\rho}{\sqrt{1+\rho^2-2\rho\cos\omega}}. \quad (49)$$

5.2. The proposed optimum word length allocation method

From Eqs. (10), (11), and (12), we can rewrite a value Δh_k as

$$\Delta h_k = 2^{-W_k}(a_k + b_k) \quad (50)$$

where

$$a_k = \sum_{j=1}^{\infty} B_{(j+W_k)} \cdot 2^{-j}$$

$$b_k = 2^{W_k}\left(\sum_{j=1}^{W_k} B_j \cdot 2^{-j} - \sum_{j=1}^{W_k} B'_j \cdot 2^{-j}\right).$$

By substituting Eq. (50) into Eq. (41), we can rewrite a variance of the errors (when quantization

is not applied) as

$$\sigma^2_{N_{TF}} = \sum_{k=0}^{14} \{\|\mathbf{S}_{Hk} \cdot \|^2 2^{-2W_k}\}. \quad (51)$$

The purpose of this paper is now summarized as follows.

$$\text{minimize} \quad \sigma^2_{N_{TF}} = \sum_{k=0}^{14} \{\|\mathbf{S}_{Hk} \cdot \|^2 2^{-2W_k}\}$$

$$\text{subject to} \quad \sum_{k=0}^{14} W_k = 15 \cdot \overline{W} \quad (52)$$

The problem is to find the optimum value of W_k for each coefficient so that effect of the truncations is minimized under a given average word length. The solution to this problem is

$$\frac{2^{W_k}}{2^{W_0}} = \frac{\|\mathbf{S}_{Hk}\|}{\|\mathbf{S}_{H0}\|}, \quad k = 0, 1, \ldots, 14. \quad (53)$$

Namely,

$$W_k = \log_2 \frac{\|\mathbf{S}_{Hk}\|}{\|\mathbf{S}_{H0}\|} + W_0, \quad k = 0, 1, \ldots, 14. \quad (54)$$

In other way,

$$W_k = \log_2 \|\mathbf{S}_{Hk}\| - \log_2 \|\overline{\mathbf{S}}_{Hk}\| + \overline{W}, \quad k = 0, 1, \ldots, 14, \quad (55)$$

where

$$\overline{W} = \frac{1}{15} \sum_{k=0}^{14} W_k, \quad (56)$$

$$\|\overline{\mathbf{S}}_{Hk}\| = \prod_{k=0}^{14} \sqrt[15]{\|\mathbf{S}_{Hk}\|}.$$

Therefore the optimum word length allocation is given by the relative SNR sensitivity SR_k as follows:

$$\Delta W_k = W_k - \overline{W} = \log_2 \frac{\|\mathbf{S}_{Hk}\|}{\|\overline{\mathbf{S}}_{Hk}\|} = \log_2 SR_k \quad (57)$$

$k = 0, 1, \ldots, 14.$

From the results in Table 1, we theoretically calculate the optimum word length allocation using Eq. (57) as illustrated in Table 2. Notice that the optimum word length allocations (ΔW_k) in Eq. (57) must be truncated into an integer value under a condition that the sum of them must equal zero.

Table 2
The theoretical optimum word length allocation (ΔW_k) based on AR(1) model

k	i	j	The optimum word length allocation (ΔW_k)			
			$\rho = 0.95$	$\rho = 0.9$	$\rho = 0.85$	$\rho = 0.8$
0		1	+1	+1	+1	+1
1	A	2	+1	+1	+1	+1
2		3	−1	−2	−2	−2
3		1	+4	+3	+3	+2
4	B	2	+4	+3	+2	+2
5		3	+5	+4	+3	+3
6		1	−4	−3	−3	−2
7	C	2	−3	−2	−1	−1
8		3	−3	−2	−1	−1
9		1	+1	+1	+1	+1
10	D	2	0	0	0	0
11		3	+2	+2	+2	+2
12		1	−2	−2	−2	−2
13	E	2	−2	−1	−1	−1
14		3	−3	−3	−3	−3

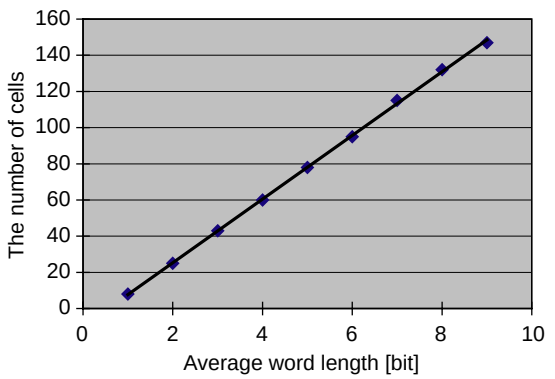

Fig. 6. A relation between an average word length and the number of APEX II cells.

In this report, we estimate the hardware volume of the circuit based on ASIC implementation with FPGA compiler II version 3.8 from VLSI Design & Education center, University of Tokyo [12]. In the case of FPGA implementation, if the number of cells for the transform is small, we can have more room for other functions in an FPGA chip. Namely, the FPGA chip can be multi-functional. Therefore, it is meaningful to reduce the number

of cells. Fig. 6 illustrates the relevance of the proposed SNR sensitivity in terms of a relation between an average word length and the number of APEX II cells. It confirms that the average word length is proportional to the number of cells.

6. Simulation results

In this section, we practically confirm an effectiveness of the optimum word length allocation by applying AR(1) model and standard images as input signals in Sections 6.1 and 6.2, respectively. In this report, we emphasize on finite-word-length effect (the different coefficients are used between encoder and decoder), so we consider an effectiveness of the proposed method in two conditions: no quantization and a small quantization step size.

6.1. Simulation results based on AR(1) model

(a) *No quantization*: The objective of this case is to verify effects of finite word length in each multiplier based on simulation results. By applying 4096-point AR(1) model as an input signal, PSNR and the SNR sensitivity $\|\mathbf{S}_{Hk}\|$ of a reconstructed signal are applied to calculate SR_k and ΔW_k using Eqs. (43) and (57), respectively. All results are concluded in Tables 3 and 4. Both theoretical analysis in Tables 1 and 2 and simulation results in Tables 3 and 4 confirm that

(1) The SNR sensitivities of different multipliers are different.

(2) The SNR sensitivity of IRTB is the most sensitive.

Table 3
Simulation results of the optimum word length allocation based on AR(1) model with $\rho = 0.95$

k	i	j	PSNR	$\|\mathbf{S}_{Hk}\|$	SR_k	ΔW_k
0		1	75.63	5.29	0.77	0
1	A	2	70.65	6.43	0.93	0
2		3	82.08	2.52	0.37	−1
3		1	53.41	68.41	9.91	3
4	B	2	58.38	26.40	3.83	2
5		3	47.60	133.45	19.34	4
6		1	78.37	2.58	0.37	−1
7	C	2	72.33	4.53	0.66	−1
8		3	72.99	4.79	0.69	0
9		1	71.73	7.22	1.05	0
10	D	2	80.83	3.34	0.48	−1
11		3	70.21	8.60	1.25	0
12		1	87.60	2.24	0.33	−2
13	E	2	80.31	3.24	0.47	−1
14		3	87.78	2.20	0.32	−2

Table 4
Simulation results of the optimum word length allocation based on AR(1) model with $\rho = 0.8$

k	i	j	PSNR	$\|\mathbf{S}_{Hk}\|$	SR_k	ΔW_k
0		1	72.19	7.87	0.75	−1
1	A	2	67.31	9.45	0.91	0
2		3	76.87	4.59	0.44	−1
3		1	53.44	68.13	6.53	3
4	B	2	58.63	25.64	2.46	1
5		3	47.55	134.33	12.88	4
6		1	73.07	4.75	0.46	−1
7	C	2	67.43	7.95	0.76	−1
8		3	68.15	8.36	0.80	0
9		1	68.33	10.67	1.02	0
10	D	2	76.76	5.34	0.51	−1
11		3	66.91	12.57	1.20	0
12		1	82.18	4.18	0.40	−1
13	E	2	74.95	6.01	0.58	−1
14		3	82.31	4.12	0.40	−1

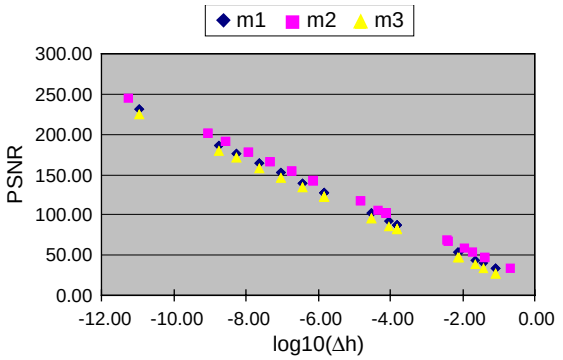

Fig. 7. Relation between PSNR and log 10(Δh) of $\mathrm{IRT_A}$ based on AR(1) with $\rho = 0.95$.

(3) The optimum word length allocation depends on frequency spectrum of an input signal.

Figs. 7 and 8 illustrate relation between PSNR and log 10(Δh) of IRT. These results confirm that we can approximate a relation between PSNR and log 10(Δh) of IRT to be a linear equation as shown in Eq. (47). Tables 5 and 6 illustrate PSNR of a reconstructed signal based on the proposed method and the existing method. The optimum word length allocations used in Tables 5 and 6 are based on the theoretical optimum word length allocations in Table 2. These results confirm an effectiveness of the proposed method since PSNR of the proposed method is significantly better than PSNR of the existing method. Notice that when the average word length equals three (in Table 5), we use a zero bit for $k = 6$. Therefore, the average word length of the proposed method equals 3.06.

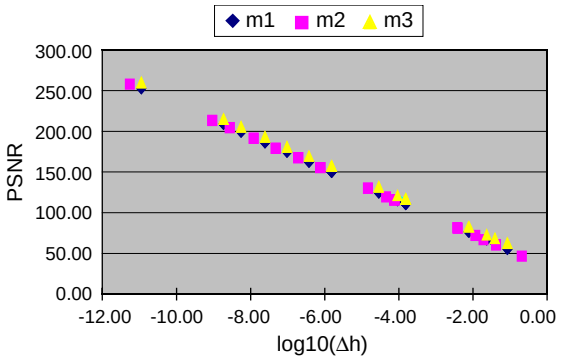

Fig. 8. Relation between PSNR and log 10(Δh) of IRTB based on AR(1) with $\rho = 0.95$.

Table 5
PSNR (dB) of a reconstructed signal based on AR(1) model with $\rho = 0.95$ and no quantization

Average word length	The existing method	The proposed method
3	30.64	44.89
4	35.20	52.32
5	45.00	60.41
6	44.46	62.80
7	69.55	73.94
8	76.12	77.14
9	77.27	83.22

Table 6
PSNR (dB) of a reconstructed signal based on AR(1) model with $\rho = 0.8$ and no quantization

Average word length	The existing method	The proposed method
3	31.78	42.48
4	36.46	52.38
5	46.16	56.80
6	45.70	61.34
7	69.05	70.06
8	75.00	74.71
9	77.81	83.84

(b) *A small quantization step size*: In this case, we confirm an effectiveness of the proposed method when a small quantization step size is applied. Table 7 illustrates PSNR of a reconstructed signal when a value of quantization step size equals two. PSNR of the proposed method is significantly better than that of the existing method. PSNRs of reconstructed signals in Table 7 are slightly worse than those in Tables 5 and 6 because those in Table 7 include quantization error and rounding error.

Table 7
PSNR of a reconstructed signal based on AR(1) model and a small quantization step size

AR(1) model	PSNR (dB)			
	The existing method		The proposed method	
	3(bit)	4(bit)	3(bit)	4(bit)
0.80	31.7	35.8	42.4	49.2
0.85	31.7	35.3	42.7	49.9
0.9	31.1	35.5	43.4	50.8
0.95	30.6	35.9	44.1	49.7

Table 8
PSNR of a reconstructed signal based on standard images and no quantization

Standard image	Bit-rate	PSNR (dB) (Different coefficients)			
		The existing method		The proposed method	
		3(bit)	4(bit)	3(bit)	4(bit)
Barbara	5.6	28.5	32.8	38.1	47.9
Lena	5.0	28.0	32.3	39.2	50.2
Couple	4.7	34.8	39.5	42.8	53.4
X-chest	6.5	25.4	29.5	37.2	46.6

6.2. Simulation results based on standard images

In this section, we applied standard images as an input signal to confirm an effectiveness of the proposed method.

(a) *No quantization*: Table 8 confirms an effectiveness of the proposed method when quantization is not applied. PSNR of the proposed method is significantly better than that of the existing method. In this case, errors are caused by only finite-word-length errors. Fig. 9 illustrates a reconstructed image based on the existing method and the proposed method. Notice that if the same coefficients are applied between

(a)

(b)

Fig. 9. A reconstructed image when quantization is not applied and average word length allocation is four: (a) A reconstructed image based on the existing method; (b) A reconstructed image based on the proposed method.

(a)

(b)

Fig. 10. A reconstructed image when a value of quantization step size equals two and average word length allocation is four: (a) A reconstructed image based on the existing method; (b) A reconstructed image based on the proposed method.

Table 9
PSNR of a reconstructed signal based on standard image and a small quantization step size

Standard image	Bit-rate	PSNR (dB) (Different coefficients)				PSNR (dB) (The same coefficients)
		The existing method		The proposed method		
		3(bit)	4(bit)	3(bit)	4(bit)	
Barbara	4.6	28.5	32.7	37.8	45.2	47.4
Lena	4.0	28.0	32.2	38.8	46.3	47.3
Couple	3.8	34.7	39.0	41.4	47.3	47.7
X-chest	5.5	25.4	29.5	36.9	44.5	47.0

encoder and decoder, PSNR equals infinity (no error).

(b) *A small quantization step size*: Table 9 confirms an effectiveness of the proposed method when a small quantization step size is applied. PSNR of the proposed method is also significantly better than that of the existing method. If the different coefficients are applied, errors are caused by all errors. On the other hand, if the same coefficients are applied, finite-word-length errors will be removed. Therefore, errors in a reconstructed image in that case depend on only rounding errors and quantization errors. From the results in Table 9, we find that finite-word-length errors dominate the other errors when an average word length is of a small value. Fig. 10 illustrates a reconstructed image based on the existing method and the proposed method.

7. Conclusion

In this report, the "SNR sensitivity" was newly defined as an indicator of how the word length truncation of multiplier coefficients affects the quality of a reconstructed image. We proposed a new word length allocation method based on the SNR sensitivity. The optimum word length allocation depends on a frequency spectrum of an input signal. Both theoretical analysis and simulation results confirm an effectiveness of the proposed method.

References

[1] M.D. Adams, F. Kossentini, Reversible integer-to-integer wavelet transform for image compression: Performance evaluation and analysis, IEEE Trans. Image Process. 9 (6) (June 2000) 1010–1024.
[2] S. Chokchaitam, M. Iwahashi, P. Zavarsky, N. Kambayashi, A bit-rate adaptive coding system based on lossless DCT, IEICE Trans. Fund. E85-A (2) (February 2002) 403–413.
[3] S. Fukuma, K. Ohyama, M. Iwahashi, N. Kambayashi, Lossless 8-point fast discrete cosine transform using lossless Hadamard transform, IEICE Technical Report, DSP99-103, October 1999, pp. 37–44.
[4] ISO/IEC 11172, Information Technology—Coding of Moving Picture And Associated Audio for Digital Storage Media at up to about 1.5m bps, MPEG Video, 1993.
[5] N.S. Jayant, P. Noll, Digital Coding of Wave Forms, Prentice-Hall, Englewood Cliffs, NJ, 1984.
[6] JPEG CD10918-1, Digital compression coding of continuous-tone still images, JPEG-9-R6, January 1991.
[7] K. Komatsu, K. Sezaki, 2D lossless discrete cosine transform, IEEE ICIP 2001, 2001, pp. 466–469.
[8] J. Reichel, G. Menegaz, M.J. Nadenau, M. Kunt, Integer wavelet transform for embedded lossy to lossless image compression, IEEE Trans. Image Compression 10 (3) (March 2001) 383–392.
[9] W. Sweldens, The lifting scheme: a construction of second generation wavelets, Technical Report 1995:6, Industrial Math. Initiative, Department of Mathematics, University of South Carolina, 1995.
[10] D. Takago, T. Takebe, Multispectral image compression using reversible WT and KLT, IEICE Trans. Fund. J84-A (3) (March 2001) 298–308.
[11] D.S. Taubman, M.W. Marcellin, JPEG 2000—Image Compression Fundamentals, Standards and Practice, Kluwer Academic Publishers, Dordrecht, 2002.
[12] VLSI Design & Education center, University of Tokyo, http://www.vdec.u-tokyo.ac.jp/English/index.html.