Electronics and Communications in Japan, Part 1, Vol. 80, No. 8, 1997
Translated from Denshi Joho Tsushin Gakkai Ronbunshi, Vol. J79-B-I, No. 8, August 1996, pp. 572–581

Bit Reduction of DCT Basis for Transform Coding

Masahiro Iwahashi and Noriyoshi Kambayashi

Faculty of Engineering, Nagaoka University of Technology, Nagaoka, Japan 940-21

Hitoshi Kiya

Faculty of Engineering, Tokyo Metropolitan University, Hachioji, Japan 192-03

SUMMARY

Extensive research has been carried out on transform coding using the discrete cosine transform (DCT) for compression of digital video data. Improvement of the video quality has been brought about by an increased number of pixels, that a reduction of the time required for coding per pixel is highly desirable. When the multiplier coefficients in the coding operation or the values of the basis coefficients of the DCT are expressed as binary numbers, it is possible to reduce the operating time and simplify the process by using as few bits as possible. It is necessary to consider the constraint that the picture quality of the reconstructed video image should not be degraded by the reduction of the number of bits. In this paper, it is shown that the number of bits of the basis coefficients can be reduced under this constraint if quantization of the transform coefficients is taken into account. Also, the limiting value of the decrease of the number of bits in specific coding examples such as JPEG is determined. Finally, it is confirmed that the quality of the reconstructed video image changes very little even if the basis coefficients are represented by fewer bits than in previous schemes, provided that the reduction is within the limit presented in this paper. © 1997 Scripta Technica, Inc. Electron Comm Jpn Pt 1, **80**(8): 81–91, 1997

Key words: Coding; cosine transform; multiplier; number of bits; hardware; video image.

1. Introduction

Recently, extensive research has been carried out on video image coding for compression of video images. Such images have a much larger amount of data than voice or document files. Among the studies, transform coding using the discrete cosine transform (DCT) [1] is the core of the international standards issued by ISO/IEC and ITU-T [2, 3, 4] and is widely studied in various areas. On the other hand, in the applications area, the use of Hi-vision, within an increased number of pixels, has progressed, so that a reduction of the time and simplification of the process of coding are required. In this paper, the coding process is simplified by expressing the multipliers in the transform coding or the values of the DCT basis in terms of as small a number of bits as possible. Several studies on the number of bits of the basis coefficients have been reported [5–7] In H.261 as a CCITT recommendation, the operation accuracy of the DCT is specified in order to avoid discrepancies of the received pictures between the decoders [2]. The operation errors generated in that process are analyzed in [5]. The number of bits needed for agreement of the reconstructed video images among different decoders and the determination procedure are reported in [6, 7]. In these reports, it is assumed that the transform coefficients that result from the sequence transform are not quantized. However, in actual coding, the number of bits of the transform coefficients is reduced by quantization and data compression is carried out. Therefore, if quantization is carried out, the number of bits of

CCC8756-6621/97/080081-11

the basis coefficients reported earlier exceeds the minimum number of bits. In this paper, with a view to simplifying the coding process, a method is studied to reduce the excess bits of the basis coefficients that occur in the conventional method. To this end, the quantization error energy is theoretically estimated as the numbers of bits for the transform coefficients and the basis coefficients are varied. It is found that the energy for the quantization error is not affected if the number of bits of the basis coefficients is smaller than the number usually used, so long as the value is above a certain limit. Therefore, the minimum number of necessary bits of the basis coefficients is obtained. If this result is used, the basis coefficients can be expressed with fewer bits than in the conventional method while the quality of the reconstructed video image remains comparable to that in the conventional method. Hence, simplification of the hardware size of the coding device and reduction of the process time become possible. In the following, the transform coding is explained in section 2. In section 3, the method for reduction of the bit size of the basis coefficients is described. In section 4, the minimum number of bits of the basis coefficients for a specific quantization is studied by simulation. Finally, in section 5, the results obtained in this paper are summarized.

2. Transform Coding [1, 8]

First the fundamental process of transform coding and the quantization of the transform coefficients are explained.

2.1. Fundamental process

The fundamental process of transform coding is explained by means of Fig. 1. First, the original signal to be coded is divided at N points into several blocks. Here, the original signal is assumed to be expressed as a binary integer with B_s bits. Next, in each block, the original signals $s(n)$ (n = 0, 1, . . . , N − 1) at N points are discrete cosine-transformed (DCT) so that N transform coefficients are obtained:

$$t(f) = I\left[T_N[s(n)]\right] = I\left[\sum_{n=0}^{N-1} b(n,f)s(n)\right] \quad (f = 0, 1, \cdots, N-1) \tag{1}$$

The values of the transform coefficients as the result of DCT are expressed as integers. The symbols $T_N[\ ]$ and $b(n, f)$ denote the DCT and the basis coefficients. The symbol $I[x]$ is the rounded-up integer closest to x. The basis coefficient $b(n, f)$ in Eq. (1) is given by

$$b(n,f) = c(f)\cos\frac{(2n+1)f\pi}{2N} \tag{2}$$

where

$$\begin{cases} c(f) = \dfrac{1}{\sqrt{N}} & (f = 0) \\ c(f) = \sqrt{\dfrac{2}{N}} & (f = 1, 2, \cdots, N-1) \end{cases}$$

Next, by quantizing the transform coefficient $t(f)$ obtained in Eq. (1), the number of bits needed in the binary representation is reduced and

$$\hat{t}(f) = Q\left[t(f)\right] \quad (f = 0, 1, \cdots, N-1) \tag{3}$$

The quantization $Q[\ ]$ is explained in detail in section 2.3. Finally, the quantized transform coefficient $\hat{t}(f)$ is inversely transformed to obtain the reconstructed signal

$$\hat{s}(n) = I\left[T_N{}^{-1}[\hat{t}(f)]\right] = I\left[\sum_{f=0}^{N-1} b(n,f)\hat{t}(f)\right] \quad (n = 0, 1, \cdots, N-1) \tag{4}$$

as the output. Here, the symbol $T_N{}^{-1}[\ ]$ indicates the inverse DCT. By taking the integer $I[\ ]$, the value of the reconstructed signal $\hat{s}(n)$ is expressed as a binary integer number with B_s bits, identical to that of the original signal $s(n)$.

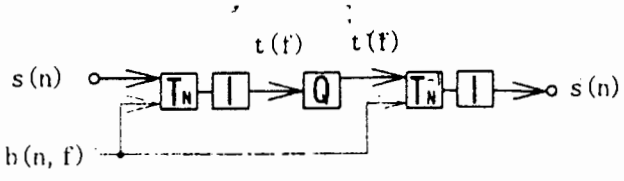

Fig. 1. Basic signal processing transform coding.

2.2. Minimum necessary bits for the basis coefficient

For the basic process in section 2.1, the basis coefficient of Eq. (2) is defined as a real number. When the

value is expressed as a binary number, an infinite number of bits is required in general. However, in the hardware realization of the coder, the word length is limited to a finite size. Therefore, the DCT operation introduces an error due to the finite word length expression of the basis coefficients. In CCITT Recommendation H.261, the word length necessary for the basis coefficients in the inverse transform in the method shown in Fig. 2 is specified. For the estimation method in the figure, the allowable ranges are specified for basis coefficients in the inverse transform of either infinite or finite bit length [2]. Here, the infinite bit length implies the 64-bit floating point representation. As a result, more than 10 bits are considered necessary for the basis coefficients [5–7]. Here, no quantization is applied to the transform coefficient.

In actual coding, the number of bits of the transform coefficient is reduced by quantization Eq. (3) for data compression. Accordingly, the minimum necessary bit size of the transform coefficient is reduced. The number of bits of the basis coefficients reported to date is considered excessive with respect to the number of bits actually needed. The objective of the present paper is to reduce the excess bits of the basis coefficients caused by the quantization of the transform coefficients and to simplify the coding process by reduction of the word length of the multiplier coefficients. In the following, the quantization of the transform coefficients is explained, after which the method for determining the minimum necessary bit size of the basis coefficients is described in section 3.

2.3. Quantization of the transform coefficients

The quantization $Q[\ ]$ in Eq. (3) is carried out with a step size $q_t(f)$ by

$$Q\left[t(f)\right] = I\left[\frac{t(f)}{q_t(f)}\right] q_t(f) \quad (f = 0, 1, \cdots, N-1) \tag{5}$$

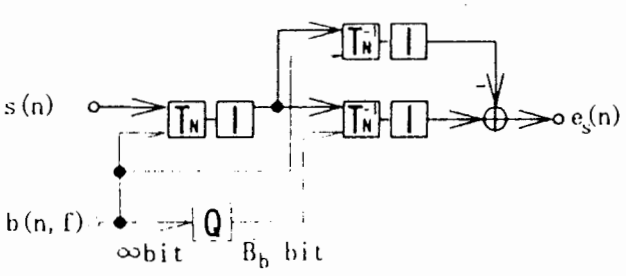

Fig. 2. Conventional method of estimation.

In the above, the value of $t(f)$ is quantized to an integer multiple of $q_t(f)$. In the transform coding, the integer value becomes the compressed information sent out from the encoder. At the decoder, the step size is multiplied by this value so that the transform coefficient is recovered. If the maximum absolute value of $t(f)$ for all the blocks is $R_t(f)$, the binary expression of the transform coefficient $t(f)$ of the f-th band requires binary codes with

$$B_t(f) = \log_2 \frac{R_t(f)}{q_t(f)} + 1 \quad (f = 0, 1, \cdots, N-1) \tag{6}$$

including a coding bit with length one. However, since the value of $B_t(f)$ in Eq. (6) is a real number in general, a binary code with $\lfloor B_t(f) \rfloor + 1$ bits is prepared in an actual hardware configuration, where $\lfloor x \rfloor$ is the maximum integer not exceeding x. On the other hand, since the energy of the signal with the value range of $[-R, R]$ is proportional to R_2 [8], the energy of the transform coefficient

$$E_t(f) = \left\{t(f)\right\}^2 \quad (f = 0, 1, \cdots, N-1) \tag{7}$$

is expressed in terms of $R_t(f)$ as

$$E_t(f) = \alpha_1 {R_t}^2(f) \quad (f = 0, 1, \cdots, N-1) \tag{8}$$

where α_1 is a proportionality constant and is dependent on the distribution function of the value of $t(f)$ [8]. When $R_t(f)$ is eliminated from Eqs. (6) and (8), one obtains

$$q_t(f) = \sqrt{\frac{4E_t(f)}{\alpha_1}} \cdot 2^{-B_t(f)} \quad (f = 0, 1, \cdots, N-1) \tag{9}$$

and hence the relationship between the step size $q_t(f)$ and the number of bits $B_t(f)$ is obtained for each bandwidth. As above, the quantization of the transform coefficient in each bandwidth is expressed in terms of the step size $q_t(f)$ or the number of bits $B_t(f)$.

2.4. Step size of transform coefficient [8]

The number of bits $B_t(f)$ for each bandwidth in Eq. (9) can be determined such that for a signal quantization error

$$e_s(n) = \hat{s}(n) - s(n) \tag{10}$$

the average energy per sample

$$E_{es} = \frac{1}{N}\sum_{n=0}^{N-1} {e_s}^2(n) \qquad (11)$$

is minimized [8]. In the following, E_{es} defined in Eq. (11) is called the error energy. A method for the determination of $B_t(f)$ that minimizes this error energy is summarized below. The energy of the quantization error of the transform coefficient

$$e_t(f) = \hat{t}(f) - t(f) \quad (f = 0, 1, \cdots, N-1) \qquad (12)$$

namely,

$$E_{et}(f) = e_t^2(f) \quad (f = 0, 1, \cdots, N-1) \qquad (13)$$

is used. Then, the error energy E_{es} is expressed as

$$E_{es} = \frac{1}{N}\sum_{f=0}^{N-1} E_{et}(f) \qquad (14)$$

when the basis coefficient is given by Eq. (2). Note that $E_{et}(f)$ in Eq. (14) is expressed in terms of the step size $q_t(f)$ [8] as

$$E_{et}(f) = \frac{1}{12}{q_t}^2(f) \qquad (15)$$

Therefore, if $q_t(f)$ is eliminated from Eqs. (9) and (15),

$$E_{et}(f) = \frac{1}{3\alpha_1}E_t(f)2^{-2B_t(f)} \qquad (16)$$

Further, if $E_{et}(f)$ is eliminated from Eqs. (14) and (16),

$$E_{es} = \frac{1}{3N\alpha_1}\sum_{f=0}^{N-1} E_t(f)2^{-2B_t(f)} \qquad (17)$$

Hence, E_{es} is expressed in terms of the number of bits $B_t(f)$ in each bandwidth. Also, the average value of $B_t(f)$ over the entire bandwidth

$$\overline{B}_t = \frac{1}{N}\sum_{f=0}^{N-1} B_t(f) \qquad (18)$$

is given beforehand. Hence, the desired value of $B_t(f)$ is the solution to the problem of determining $B_t(f)$ that minimizes the error energy in Eq. (17) under constraint of Eq. (18). This solution can be obtained by the Lagrange method for the undetermined multiplier. As a result, when the number of bits $B_t(f)$ of each bandwidth is

$$B_t(f) = \overline{B}_t + \frac{1}{2}\log_2 \frac{E_t(f)}{\left\{\prod_{g=0}^{N-1} E_t(g)\right\}^{1/N}} \qquad (19)$$

the value of E_{es} is a minimum [8]. When Eq. (19) is substituted into Eq. (9), we obtain

$$\begin{aligned} q_t(f) &= \alpha_t \cdot 2^{-\overline{B}_t} \\ &= q_t \quad (f = 0, 1, \cdots, N-1) \end{aligned} \qquad (20)$$

where

$$\alpha_t = \frac{1}{\sqrt{\alpha_1}}\left\{\prod_{g=0}^{N-1} E_t(g)\right\}^{1/2N}$$

Hence, $q_t(f)$ for the minimum E_{es} is given in terms of $\overline{B}_t$. In this case. it is found from Eq. (20) that $q_t(f)$ is identical over the entire bandwidth.

In the above, we describe the quantization of the transform coefficient, with step size $q_t(f)$ for each bandwidth as a parameter, as shown in Eq. (5). Also, when the error energy in Eq. (11) is minimized, $q_t(f)$ is expressed by the average number of bits $\overline{B}_t$ and has an identical value over the entire bandwidth.

3. Method of Reducing the Bit Size of the Basis Coefficients

Next, we describe a method of making the basis coefficients have a finite word length when the transform coefficients are quantized as explained in section 2.3.

3.1. Estimation method

As described in section 2.2, this paper investigates the minimum necessary number of bits for the basis coefficients when the transform coefficients are quantized, within the range where the use of a finite word length in the basis coefficients does not affect the error energy E_{es} in Eq. (11). To this end, in place of the conventional estimation method shown in Fig. 2, we use that in Fig. 3, in which the finite word length in the basis coefficient

$$\hat{b}(n,f) = Q\left[b(n,f)\right] \tag{21}$$

is introduced. Therefore, in place of Eq. (4), the basis coefficient $\hat{b}(n,f)$ with a limited word length is used. The inverse transform is carried out by

$$\begin{aligned}\hat{s}(n) &= I\left[T_N{}^{-1}[\hat{t}(f)]\right] \\ &= I\left[\sum_{f=0}^{N-1}\hat{b}(n,f)\hat{t}(f)\right] \\ &\qquad (n = 0, 1, \cdots, N-1)\end{aligned} \tag{22}$$

3.2. Finite word length for the basis coefficient

In this paper, the word length of the basis coefficient $b(n,f)$ in Eq. (21) is made finite by means of the step size $q_b(f)$ as given in Eq. (5), such that

$$Q\left[b(n,f)\right] = C_b(f)q_b(f) \tag{23}$$

where

$$\begin{aligned}C_b(f) &= I\left[\frac{b(n,f)}{q_b(f)}\right] \\ &\qquad (n = 0, 1, \cdots, N-1,\ f = 0, 1, \cdots, N-1)\end{aligned}$$

Here, using a positive integer $Q_b(f)$, $q_b(f)$ is

$$q_b(f) = 2^{-Q_b(f)} \quad (Q_b(f) \text{ is a positive integer}) \tag{24}$$

By Eq. (23), the value of $b(n,f)$ originally defined by an infinite word length is now approximated by an integer multiple of $q_b(f)$. Hence, the multiplication of $\hat{b}(n,f)$ and $\hat{t}(f)$ is carried out by first multiplying the integer $\hat{t}(f)$ and the integer $C_b(f)$, then multiplying by $q_b(f)$. Since $Q_b(f)$ is an integer, the multiplication becomes a shift operation of binary numbers. If the maximum of the absolute value of $b(n,f)$ for all n's in each bandwidth f is $R_b(f)$, the binary representation of the integer $C_b(f)$ requires binary codes with lengths

$$\begin{aligned}B_b(f) &= \log_2 \frac{R_b(f)}{q_b(f)} + 1 \\ &= -\log_2 q_b(f) + \log_2 R_b(f) + 1 \\ &\qquad (f = 0, 1, \cdots, N-1)\end{aligned} \tag{25}$$

bits, including the coding bit. The value of $R_b(f)$ can be computed from the maximum absolute value of Eq. (2). Since the value of $B_b(f)$ in Eq. (25) is a real number in general, a binary signal with $\lfloor B_b(f) \rfloor + 1$ bits is prepared when the hardware is actually configured.

In the above, the finite word length representation of the basis coefficient was described by Eq. (23) with a parameter $q_b(f)$. It is seen from Eqs. (24) and (25) that the number of bits $B_b(f)$ needed in the binary expression of the basis coefficient $\hat{b}(n,f)$ with a finite word length is proportional to $Q_b(f)$.

3.3. Step size of the basis coefficient

As explained in section 2.4, the step sizes of the transform coefficient minimizing the error energy became equal in all bandwidths. The step size can then be expressed by the average number of bits. Hence, for the case where the step sizes of the basis coefficient are equal in all bands,

$$q_b(f) = q_b \quad (f = 0, 1, \cdots, N-1) \tag{26}$$

let us study the relationship between the step size and the average number of bits. The average number of bits $\overline{B}_b$ of the basis coefficient is defined as

$$\overline{B}_b = \frac{1}{N}\sum_{f=0}^{N-1} B_b(f) \tag{27}$$

When Eqs. (25) and (26) are substituted into the above,

$$\overline{B}_b = -\log_2 q_b + \log_2 2\left\{\prod_{g=0}^{N-1} R_b(g)\right\}^{1/N} \tag{28}$$

When the above is modified, one obtains

$$q_b = \alpha_b 2^{-B_b} \quad (f = 0, 1, \cdots, N-1) \tag{29}$$

where

$$\alpha_b = 2\left\{\prod_{g=0}^{N-1} R_b(g)\right\}^{1/N}$$

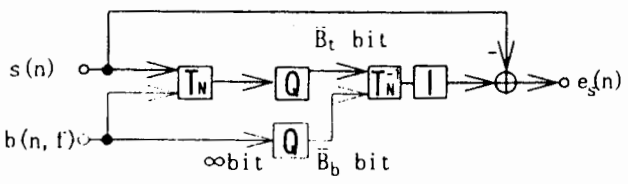

Fig. 3. Proposed method of estimation.

Therefore, the relationship is obtained between the step size q_b and the average number of bits $\overline{B}_b$ when the word length of the basis coefficient is made finite by using step sizes with the same values in all bands.

3.4. Effect of finite word length on error energy

In sections 2.3 and 2.4, the quantization of the transform coefficient was described. In sections 3.2 and 3.3, the use of finite word length of the basis coefficient is described in terms of the average number of bits $\overline{B}_t$, and $\overline{B}_b$. In what follows, the effects of these processes on the error energy are analyzed theoretically. In this paper, the final objective is to derive the minimum $\overline{B}_b$ for which the value of E_{es} is about the same as that for $\overline{B}_b = \infty$ with a given $\overline{B}_t$.

3.4.1. Components of quantization error

First, let us investigate the components contained in the quantization error $e_s(n)$ defined by Eq. (10). To simplify the expression, the forward transform $T_N[x]$ using $b(n, f)$ is written as $b(n, f) * x$. Then, the transform coefficient is

$$t(f) = b(n, f) * s(n) \tag{30}$$

Similarly, if the inverse transform $T_N^{-1}[x]$ using $\hat{b}(n, f)$ with a finite word length is expressed as $\hat{b}(n, f) * x$, the reconstructed signal is

$$\hat{s}(n) = \hat{b}(n, f) * \hat{t}(f) + e_i \tag{31}$$

where

$$e_i = I\left[\hat{b}(n, f) * \hat{t}(f)\right] - \hat{b}(n, f) * \hat{t}(f)$$

Here, e_i is the error generated by the integer operation $I[\ ]$. Further, the basis coefficient with finite word length is separated into the true value and the error:

$$\hat{b}(n, f) = b(n, f) + e_b(n, f) \tag{32}$$

Summarizing the above, one obtains from Eqs. (10), (12), (30) to (32)

$$\begin{aligned} e_s(n) = {} & e_b(n, f) * b(n, f) * s(n) \\ & + b(n, f) * e_t(f) \\ & + e_b(n, f) * e_t(f) \\ & + e_i \end{aligned} \tag{33}$$

Hence, the quantization error $e_s(n)$ is now expressed by four terms.

3.4.2. Energy of components

Let us study the energy of these four terms. We consider the case where the step sizes $q_t(f)$ and $q_b(f)$ are equal in the entire bandwidth. The case where they are different for each band will be considered in section 3.6. The values of the signals contained in each term in Eq. (33) are within the ranges listed in Table 1. For instance, since the signal $e_t(f)$ is the quantization error for the step size 2^{Q_t}, the maximum absolute value is 0.5 times the step size. Hence, the distribution range of the signal $e_t(f)$ is $\pm 2^{Q_t - 1}$. In the table, Q_t and Q_b are defined by q_t and q_b in Eqs. (20) and (26) as

$$Q_t = \log_2 q_t, \quad f = 0, 1, \cdots, N-1 \tag{34}$$

$$Q_b = -\log_2 q_b, \quad f = 0, 1, \cdots, N-1 \tag{35}$$

where Q_b is a positive integer as mentioned in Eq. (24). On the other hand, since an arbitrary positive integer is used for q_t [3, 4], Q_t is a positive real number. In the following, they are called logarithmic steps. When Eqs. (20) and (29) are substituted into Eqs. (34) and (35),

$$Q_b = \overline{B}_b - \log_2 \alpha_b \tag{36}$$

$$Q_t = -\overline{B}_t + \log_2 \alpha_t \tag{37}$$

The logarithmic steps are proportional to the average numbers of bits. When Table 1 is used, the distribution ranges of the amplitude values of the four terms in Eq. (33) can be found. The obtained results are listed in Table 2. In Table 2, the number of bits for the input signal $s(n)$ is B_s. In the forward transform by the basis coefficient $b(n, f)$, the distribution ranges of the signals are not changed. This implies that, if the forward transform is interpreted as processing by an FIR (Finite Impulse Response) filter with the transfer function

Table 1. Range of signals

quantity	range
e_t	$\pm 2^{Q_t - 1}$
s	$\pm 2^{B_s - 1}$
e_i	$\pm 2^{-1}$
e_b	$\pm 2^{-Q_b - 1}$

Table 2. Range of the signals

quantity	range
$e_b * b * s$	$\pm 2^{-Q_b + B_s - 2}$
$b * e_t$	$\pm 2^{Q_t - 1}$
$e_b * e_t$	$\pm 2^{Q_t - Q_b - 2}$
e_i	$\pm 2^{-1}$

$$B(z, f) = \sum_{n=0}^{N-1} b(n, f) z^{-n} \quad (f = 0, 1, \cdots, N-1) \tag{38}$$

the maximum gain of the filter is less than unity.

On the other hand, the energy of the signal with its value distributed within the range of $\pm 2^{-1}R$ is proportional to R^2 [8]. Therefore, the energy E_{es} of the error signal $e_s(n)$ in Eq. (33) is

$$E_{es} = \beta_1 \cdot 2^{2(-Q_b + B_s - 1)} + \beta_2 \cdot 2^{2Q_t} + \beta_3 \cdot 2^{2(Q_t - Q_b - 1)} + \beta_4) \tag{39}$$

where the coefficients β_1, β_2, β_3, and β_4 are constants dependent on the distribution functions of the signal value. It is known that they are all 12^{-1} for the uniform distribution [8].

3.4.3. Relationship between the error energy $\overline{B}_t$ and $\overline{B}_b$

The dotted lines in Fig. 4 indicate the relationship of Eq. (39) for $B_s = 8$. The error energy E_{es} is indicated by the PSNR (Peak Signal-to-Noise Ratio), defined by

$$\sigma_{es} = 10 \log_{10} \frac{(2^{B_s} - 1)^2}{E_{es}} = -10 \log_{10} E_{es} + 20 \log_{10}(2^{B_s} - 1) \tag{40}$$

From the figure, it is found that σ_{es} is proportional to Q_b if Q_b is below a certain value and that σ_{es} depends not on Q_b but on Q_t. This implies that the third and fourth terms are negligibly small for the entire Eq. (39), so that either the first or second term becomes dominant. From the relationship of the exponents of the first and the second terms, Eq. (39) can be expressed as:

(i) $-Q_b + B_s - 1 > Q_t$,

$$E_{es} \approx \beta_1 \cdot 2^{2(-Q_b + B_s - 1)} \tag{41}$$

(ii) $-Q_b + B_s - 1 < Q_t$,

$$E_{es} \approx \beta_2 \cdot 2^{2(Q_t)} \tag{42}$$

When Eqs. (36) and (37) are substituted into Eqs. (41) and (42),

(i) $\overline{B}_b < \overline{B}_t + \log_2 \frac{\alpha_b}{\alpha_t} + B_s - 1$,

$$E_{es} \approx (\beta_1 \alpha_b^2 \cdot 2^{2(B_s - 1)}) \cdot 2^{-2\overline{B}_b} \tag{43}$$

(ii) $\overline{B}_b > \overline{B}_t + \log_2 \frac{\alpha_b}{\alpha_t} + B_s - 1$,

$$E_{es} \approx (\beta_2 \alpha_t^2) \cdot 2^{-2\overline{B}_t} \tag{44}$$

As above, the error energy E_{es} is expressed as a function of the average number of bits $\overline{B}_t$ and $\overline{B}_b$.

3.5. Method of determination of the minimum number of bits of basis coefficient

As the results of the investigation up to this point, it is found that the minimum value of $\overline{B}_b$ under a given $\overline{B}_t$ is the value of $\overline{B}_b$ when the first term in Eq. (39) is negligible in comparison to the second term. Hence, by using an extremely small number ε, the ratio of both is written as

$$\varepsilon = \frac{\beta_2 2^{2(-Q_{b,\min} + B_s - 1)}}{\beta_1 2^{2Q_t}} \tag{45}$$

When the above is modified,

$$Q_{b,\min} = -Q_t + B_s - 1 + \frac{1}{2} \log_2 \frac{\beta_2}{\beta_1 \varepsilon} \tag{46}$$

Further, by substituting Eqs. (36) and (37),

$$\overline{B}_{b,\min} = \overline{B}_t + B_s - 1 + \frac{1}{2} \log_2 \frac{\alpha_b^2 \beta_2}{\alpha_t^2 \beta_1 \varepsilon} \tag{47}$$

Now the estimation equation for the minimum number of bits of the basis coefficient is obtained. It is found that $\overline{B}_{b,\min}$ is proportional to $\overline{B}_t$ and B_s. Hence, the minimum number of bits of the basis coefficient is clear theoretically in the case where the quantization of the transform coefficient is taken into account.

The number of bits required for the basis coefficient has been computed under the assumption that quantization of the transform coefficient is not carried out [2, 5–7]. However, in

actual coding, data compression is performed by quantizing the transform coefficient. The minimum number of bits needed in this case is now given by Eq. (47) in this paper. Hence, the number of bits reported previously exceeds the minimum necessary number of bits given by Eq. (47). By eliminating these unnecessary bits, the coding process can be simplified. The minimum number of bits in a specific coding example is computed in section 4.

3.6. Case with different step sizes in each band

As a result of minimization of the error energy, the step size $q_t(f)$ of the transform coefficient remains unchanged for all bandwidths. However, in coding schemes such as JPEG and MPEG, the step sizes are different for each band [3, 4]. This is due to consideration of human vision characteristics. In such cases, the transform coefficients only in the f-th band out of N bands are quantized by the given step size $q_t(f)$. Then, the minimum logarithmic step of the basis coefficients is

$$Q_{b,\min}(f) = -Q_t(f) + B_s - 1 + \frac{1}{2}\log_2 \frac{\beta_2(f)}{\beta_1(f)\varepsilon} \tag{48}$$

as in Eq. (46). Also, from the definition of the logarithmic step,

$$Q_b(f) = -\log_2 q_b(f) \tag{49}$$

Hence, if $Q_b(f)$ is eliminated from Eqs. (48) and (49), the corresponding step size $q_b(f)$ can be obtained. When this is substituted into Eq. (25), the minimum number of bits of the basis coefficients for each band can be obtained.

3.7. Application to two-dimensional signal processing

The discussions above are intended for one-dimensional signal processing. In order to apply the present method to two-dimensional signal processing, it is necessary to replace $T_N[\ ]$ and $T_N^{-1}[\ ]$ with the following:

$$T_N\left[s(n_1,n_2)\right] = \sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1} b(n_1,n_2,f_1,f_2)\ s(n_1,n_2)$$
$$(f_1 = 0,1,\cdots,N-1,\ f_2 = 0,1,\cdots,N-1) \tag{50}$$

and

$$T_N^{-1}\left[\hat{t}(n_1,n_2)\right] = \sum_{f_1=0}^{N-1}\sum_{f_2=0}^{N-1} b(n_1,n_2,f_1,f_2)\ \hat{t}(f_1,f_2) \tag{51}$$
$$(n_1 = 0,1,\cdots,N-1,\ n_2 = 0,1,\cdots,N-1)$$

Here, in place of Eq. (2), the following is used for the basis coefficient $b(n_1, n_2, f_1, f_2)$:

$$b(n_1,n_2,f_1,f_2) = c(f_1,f_2)\cos\frac{(2n_1+1)f_1}{2N}\pi\ \cos\frac{(2n_2+1)f_2}{2N}\pi \tag{52}$$

where

$$c(f_1,f_2) = c(f_1)c(f_2)$$

4. Simulation

In the following, the effectiveness of the method described in section 3 is confirmed by computer simulation.

4.1. Signal model

First, an experiment is carried out by using a signal model approximating the characteristics of the video image data. Here, let us use the output from the transfer function

$$H(z) = \frac{1}{1-\rho z^{-1}} \tag{53}$$

with a random number as the input. The value of the output signal is assumed to be an integer of B_s bits. Also, ρ is the correlation factor of the signal, indicating that the energy is deviated toward the lower frequency region if the value is closer to 1.

The solid lines in Fig. 4 indicate the results of the error energy σ_{es} versus the log step. Here, B_s = 8 and ρ = 0.9. Also, the dotted lines in the same figure indicate the theoretical values from Eq. (39). A uniform distribution is assumed. As shown in Eq. (39), the theoretical equation for the error energy is determined by the value range of the

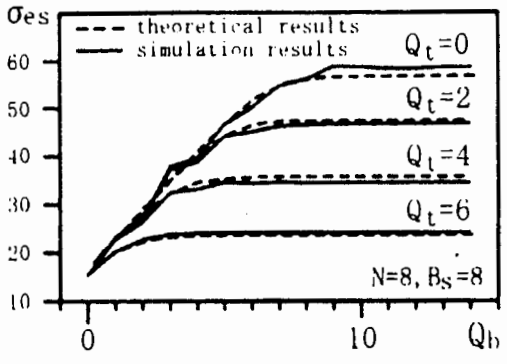

Fig. 4. Relation between error energy and log-step.

signal and does not depend on the correlation factor ρ. From the figure, it is confirmed that the simulation results agree well with the theoretical values. If the minimum $Q_{b,\min}$ for which the deviation from σ_{es} for $Q_b = 14$ is less than δ dB is read from Fig. 4, the results indicated with circles in Fig. 5 are obtained. The solid line in this figure indicates the theoretical values of Eq. (46) that agree well with the measured values. Here, $\beta_1 = \beta_2$ in Eq. (46). Also, the value of infinitesimal number ε is 2^{-4}. Then, from

$$\delta = 10 \log_{10}(1 + \varepsilon) \tag{54}$$

one finds $\delta = 0.3$ dB. Subjectively, such a difference cannot be detected. Also, Fig. 6 shows the relationship between the minimum number of bits $\overline{B}_{b,\min}$ of the basis coefficient and PSNR σ_{es}. The minimum number of bits $\overline{B}_{b,\min}$ is computed from the minimum log step $Q_{b,\min}$ in Fig. 5 by way of Eq. (36). For instance, if the error energy is 32 dB, the energy of the quantization error contained in the reconstructed signal is not affected even when the number of bits of the basis coefficient is 3 instead of more than 10 as required in the conventional method.

4.2. Video image data

Next, the experimental results are shown in the case where the standard video image is used as the input. Figure 7 shows the relationship between the error energy and the log step. As the video image, one frame of the interlaced dynamic picture "flower garden" was used. No field separation was invoked. For the quantization of the transform coefficient, step sizes different for each band as shown in [3] are used after scaler multiplication. $\overline{Q}_b$ is the average value of $Q_b(f)$. From the figure, it is found, as in Fig. 4, that σ_{es} is almost proportional to $\overline{Q}_b$ while σ_{es} no longer changes if $\overline{Q}_b$ is more than a certain value. In Fig. 8, the value of $\overline{Q}_b$ at this boundary, namely

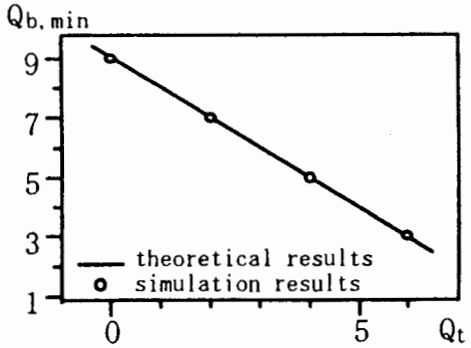

Fig. 5. Relation between error energy and log-step.

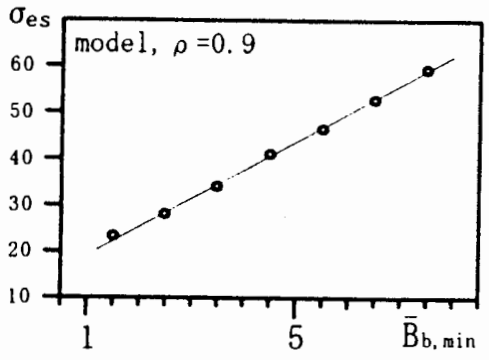

Fig. 6. Relation between error energy and minimum number of bits.

$\overline{Q}_{b,\min}$ and σ_{es} are plotted. From the figure, it is found that $\overline{B}_{b,\min} = 5$ bits if $\sigma_{es} = 31$ dB. This implies that the quality of the decoded picture represented with basis coefficients of 5 bits on the average does not degrade in comparison with more bits as long as the image is of 31 dB quality. In Fig. 9, the decoded images for $\sigma_{es} = 31$ dB are shown with $\overline{B}_b = 5$ and $\overline{B}_b = 14$. Comparison shows that there is little difference in the image quality.

Previously, it has been considered necessary that the number of bits of the basis coefficient be more than 10 [3, 4]. However, the minimum bit lengths of the basis coefficient corresponding to the image quality of the decoded image data in Fig. 8 indicate that the image quality is not affected by reduction of $\overline{B}_b$ to 5 bits if σ_{es} is 31 dB. In the practical case of σ_{es} between 30 and 40 dB, it is sufficient to use fewer bits than the number reported for the basis coefficient.

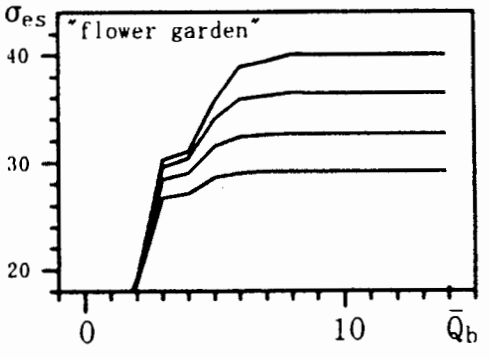

Fig. 7. Relation between error energy and log-step.

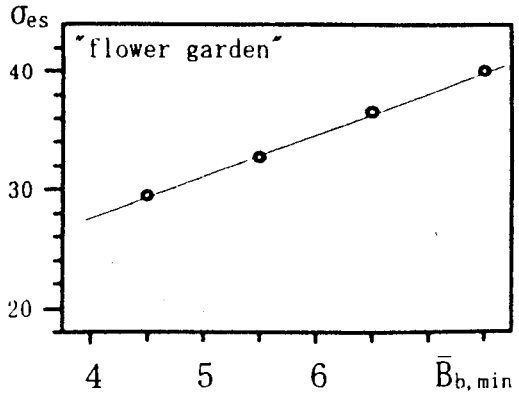

Fig. 8. Relation between error energy and minimum number of bits.

(a) $\overline{B}_b = 5$ bit

(b) $\overline{B}_b = 14$ bit

Fig. 9. Decoded image data (interlaced).

5. Conclusions

In this paper, in order to simplify the image coding process, we propose a method of representing the basis coefficient values in the transform coding using DCT by as small a bit length as possible. First, the relationship between the bit size of the basis coefficient, the transform coefficient and the energy of the quantization error was theoretically analyzed. By means of this, it is possible to theoretically estimate the image degradation of the decoded image data due to the bit size of the basis coefficient when the transform coefficient is quantized. Next, based on the results obtained the minimum necessary number of bits required for the basis coefficient when the transform coefficient is quantized was derived. Further, the validity of the theoretical values was confirmed by simulation. Our investigation was confirms that the previously reported number of bits exceeds the minimum necessary number of bits if the transform coefficient is quantized and that the quality of the decoded image data is little affected if these excess bits are removed.

In the present estimation, a standard algorithm is used in the DCT operation. In the future, studies will be conducted for cases in which various high speed algorithms are used for DCT operations.

Acknowledgments. The authors thank Mr. K. Ohyama, chief researcher at Graphics Communication Laboratories, Ltd., for his cooperation with our research.

REFERENCES

1. K.R. Rao and P. Yip. Discrete Cosine Transform — Algorithms, Advantages, Applications. New York: Academic 1990. (H. Yasuda and H. Fujiwara tr.), Image Coding Technologies—DCT and Its International Standard. Ohm Press (1992).
2. CCITT Recommendation H.261. Video CODEC for Audiovisual Devices at px 64 kbit/s (Dec. 1990).
3. JPEG CD10918-1. Digital compression coding of continuous-tone still images. JPEG-9-R6 (Jan. 1991).
4. MPEG CD 11172. Coding of moving picture and associated audio for digital strage media at up to about 1.5 Mbps. MPEG VIDEO COMMITTEE (Dec. 1990).
5. T. Mochizuki. Operation error analysis of inverse DCT finite word length operation. Trans. I.E.I.C.E. (A), **J77,** No. 3, pp. 357–368 (March 1994).
6. Y. Kikuchi and M. Akamine. Perfectly decodable DCT operation accuracy in video image coding using

DCT. Trans. I.E.I.C.E. (A), **J74,** No. 7, pp. 1116–1120 (July 1991).
7. T. Sunagawa, H. Ochi, and S. Kinjo. Determination of perfectly decodable operation accuracy in variable block size DCT. Tech. Rep. I.E.I.C.E., **CAS94–85** (Jan. 1995).
8. A.N. Akansu and R.A. Haddad. Multiresolution Signal Decompository—Transforms, Subbands, Wavelets. New York, Academic (1992).

AUTHORS (from left to right)



Masahiro Iwahashi (member) graduated from Tokyo Metropolitan University, Department of Electrical Engineering, in 1988 and completed one M.S. course in 1990. In that year, he joined Shin Nippon Steel, Ltd., Electronics Research Laboratory. In 1991, he was on temporary assignment at G C Technology, Ltd., where he was engaged in research on digital motion picture coding. In 1993, he became a research associate at Nagaoka University of Technology. He has been engaged in research on digital signal processing, particularly on high efficiency coding of image data. He is a member of IEEE.

Noriyoshi Kambayashi (member) graduated from Shinshu University, Department of Communication Engineering, in 1963 and became a research associate at Tokyo Institute of Technology, Department of Electron Physics, in 1967. In 1978, he became an associate professor at Nagaoka University of Technology, where he is now a professor. He has been engaged in research on electron circuits and digital signal processing and its applications. His publications include "Filter Theory and Design" and "Basic Circuit Engineering" (coauthored).

Hitoshi Kiya (member) graduated from Nagaoka University of Technology, Department of Electrical and Electronic Systems in 1980 and completed the M.S. course in 1982. In that year, he became a research associate at Tokyo Metropolitan University, Department of Electrical Engineering. He is presently an associate professor in Department of Electronic and Information Engineering. He holds a D.Eng. degree. He has been engaged in research on digital signal processing, especially multirate signal processing and its applications. His publications include "Fast Fourier Transform and Applications," "Introduction to Digital Signal Processing Techniques" and "Multirate Signal Processing." He is a member of the Image Electronics Society and IEEE.