

Word Length Assignment of DC Lossless DWT

Masahiro Iwahashi* and Hitoshi Kiya†

*Nagaoka University of Technology, Niigata, 980-2188, Japan

†Tokyo Metropolitan University, Tokyo, 191-0065, Japan

Abstract— This report investigates the minimum word length of coefficients and signals such that a lifting structured discrete wavelet transform (DWT) becomes lossless for a constant valued input signal (DC signal). It guarantees that the output signal contains no error in spite of rounding of coefficients and signals. This DC lossless condition is a necessary condition for the regularity which has been analyzed to improve performance of a transform. When the regularity is not satisfied, the DWT has checker board artifact observed in a reconstructed signal and DC leakage which decreases the coding gain. We derive the condition utilizing mutual affect between rounding of signals and that of coefficients. It is confirmed that the minimum word length under the condition derived by our analysis is shorter than that determined by a conventional analysis. It contributes to build a low complexity DWT under the condition.

I. INTRODUCTION

In this report, we investigate the minimum word length of coefficient values and that of signal values such that the DWT becomes lossless for a constant valued input signal (DC signal). Under this DC lossless condition, it is guaranteed that the output signal contains no error in spite of rounding of coefficients and signals inside the DWT.

In case of the 5-3 DWT in JPEG 2000 for lossless coding, benefiting from its lifting structure [1-3], lossless reconstruction of any signal is guaranteed even though signals and coefficients are rounded. On the contrary, it does not hold good for the 9-7 DWT because of scaling for adjustment of DC gain of a low pass filter in the forward DWT [4]. However, we have pointed out that it became possible to be lossless for a DC signal under a certain condition on word length of coefficients and signals [5].

This DC lossless condition is a necessary condition for the regularity which has been analyzed by numerous researchers to improve coding performance of a transform [6-8]. When the regularity is not satisfied, the DWT has the checker board artifact which is observed in a reconstructed signal as unnecessary high frequency noise [6]. It also brings about the DC leakage which decreases the coding gain.

The regularity has been structurally guaranteed for a quadrature mirror filter bank [6], a bi-orthogonal filter bank [7] and the DCT [8] respectively. However, since these previous methods are based on the lattice structure, these are not directly applicable to the lifting structure of the 9-7 DWT. Beside these relations to the regularity, the DC lossless condition itself is also considered to be useful for white balancing of a video system in which the DC signal is used as a test input for calibration [9].

This report investigates the minimum word length of the 9-7 DWT under the DC lossless condition. We have already

analyzed the DC lossless condition utilizing mutual effect between rounding of signals and that of coefficients [10]. In this report, we assign the optimum word length to each coefficient, introducing tolerance for errors as a parameter to simultaneously control both of word length of signals and that of coefficients.

II. WORD LENGTH AND DC LOSSLESS

A. Word Length and Rounding Error

Assuming the fixed point binary representation of signal values and coefficient values [4], we use the rounding operation defined as

$$\begin{cases} R_0[s] = \lfloor s' \rfloor = s' - (s' \bmod 1), & s' = s + 2^{-1}, \\ R_{F_s}[s] = R_0[s2^{F_s}]2^{-F_s}, & 0 \leq F_s \in \mathbf{Z}. \end{cases} \quad (1)$$

The operation $R_{F_s}[s]$ shortens the fraction part of word length of a value s into F_s [bit]. It also generates the error:

$$\Delta_{F_s}[s] = s - R_{F_s}[s]. \quad (2)$$

Denoting the integer part as I_s [bit], total word length W_s [bit] of a signal s is defined as

$$W_s = I_s + F_s + 1 \quad (3)$$

including 1 [bit] for the sign part. Similarly, total word length W_c [bit] of a coefficient c is defined as

$$W_c = I_c + F_c + 1. \quad (4)$$

Especially, we utilize properties of the rounding operation and its errors [10,11]:

$$\left| \Delta_{F_s}[s] \right| \leq 2^{-1-F_s} \quad (5)$$

$$R_{F_s}[s]2^{F_s} = p \in \mathbf{Z} \quad (6)$$

$$R_{F_s}[s]2^{F_s} = p \Leftrightarrow |s| \leq (p + 2^{-1})2^{-F_s} \quad (7)$$

$$s2^{F_s} \in \mathbf{Z} \Rightarrow R_{F_s}[s+t] = s + R_{F_s}[t] \quad (8)$$

which hold for any real number s and t . In Eq.(6) and (7), p is an integer determined by s under a given integer F_s .

B. Definition of DC Lossless

A forward DWT (two channel 1D filter bank) splits an input signal $x(n)$ with F_x [bit] fraction part into a low band signal $y_1(m)$ and a high band signal $y_2(m)$. Fraction part of these band signals are rounded to F_b [bit] at output. These are

synthesized by the backward DWT and rounded to F_X [bit] at output to reconstruct a signal $w(n)$. Inside the DWT, signals are multiplied by coefficients with F_C [bit] and rounded into F_S [bit]. Our concern is to find the minimum of F_C and F_S for a given F_X such that the 9-7 DWT becomes DC lossless.

The DC lossless is defined as the conjunction of the following two propositions:

$$\forall n \in \mathbb{N}, m \in \mathbb{M} (x(n) = d \rightarrow y_2(m) = 0) \quad (9)$$

$$\forall n \in \mathbb{N} (x(n) = d \rightarrow w(n) = d) \quad (10)$$

for a given constant value d with F_X [bit] fraction part. When the proposition in Eq.(9) holds, the DWT has no DC leakage for the DC input signal with value d . Similarly, when the proposition in Eq.(10) is true, the reconstructed signal $w(n)$ contains no checker board artifact for the DC input signal.

III. DC LOSSLESS CONDITION ON WORD LENGTH

A. Model for Error Analysis

Fig.1(a) illustrates a multiplier in the DWT circuit. An input value s has F_S [bit] fraction part and multiplied by a coefficient c' . The coefficient is originally designed as a real number c . It is rounded to a rational number c' in implementation. Fig.3(b) illustrates our model for theoretical analysis. The output value is described as

$$\begin{cases} s' = cs + e \\ e = R_{F_S} [\Delta_{F_S} [cs] - \Delta_{F_C} [c]s] - \Delta_{F_S} [cs] \end{cases} \quad (11)$$

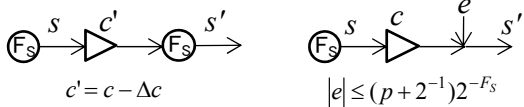
In this model, both of coefficient error and signal error are unified to the error e . Utilizing Eq.(5) and (7), its maximum value is described as

$$|e| \leq (p + 2^{-1})2^{-F_S} \quad (12)$$

Note that the parameter p to control word length of a coefficient c is included. It is equivalent to

$$2^{-F_C + F_S + I_S - 1} \leq p \quad (13)$$

From this inequality, it becomes possible to consider mutual effect of the coefficient error and the signal error [10].



(a) Multiplier. (b) Model for analysis.

Fig. 1 A multiplier and its models for analysis.

B. DC Equivalent Circuit

When the input signal is restricted to a DC signal, $x(n)$ can be described as a scalar x independent of n . The delay z^{-1} can be treated as 1 and $(1+z^{-1})$ can be replaced by 2. As a result, we can use equivalent circuits for a DC input signal in Fig.2 to derive the condition [10].

In Fig.2(a), a scalar x with F_X [bit] fraction part is multiplied by a rational number c_i , $i \in I$ and rounded to F_S [bit]. The signals are rounded to F_B [bit] at its output to produce two scalars $[y_1 \ y_2]$. The unified errors inside the circuit are described as

$$\begin{cases} e_i = R_{F_S} [\Delta_{F_S} [c_i s_i] - \Delta_{F_C} [c_i] s_i] - \Delta_{F_S} [c_i s_i] \\ f_i = R_{F_S} [\Delta_{F_S} [c_i t_i] - \Delta_{F_C} [c_i] t_i] - \Delta_{F_S} [c_i t_i] + f_i'' \end{cases} \quad (14)$$

where

$$\begin{cases} \begin{bmatrix} s_1 & s_3 & s_5 \\ s_2 & s_4 & s_6 \end{bmatrix} = \begin{bmatrix} 2x & 2(x+s'_2) & s_4 \\ 2(x+s'_1) & 2(s_2+s'_3) & s_3+s'_4 \end{bmatrix} \\ \begin{bmatrix} t_1 & t_3 & t_5 \\ t_2 & t_4 & t_6 \end{bmatrix} = \begin{bmatrix} 2(t_3-t'_2) & 2(t'_5-t'_4) & y_1 \\ 2(t_4-t'_3) & 2t'_6 & y_2 \end{bmatrix} \end{cases}$$

and

$$\begin{cases} s'_i = c_i s_i + e_i = R_{F_S} [c_i s_i] \\ t'_i = c_i t_i + f_i = R_{F_S} [c_i t_i] \end{cases}$$

Similarly to Eq.(12), these errors are described with the parameters p_i and q_i to control word length of coefficients as

$$\begin{cases} |e_i| \leq (p_i + 2^{-1})2^{-F_S} \\ |f_i| \leq (q_i + 2^{-1})2^{-F_S} \end{cases} \quad (15)$$

for a given word length F_S [bit] of signals.

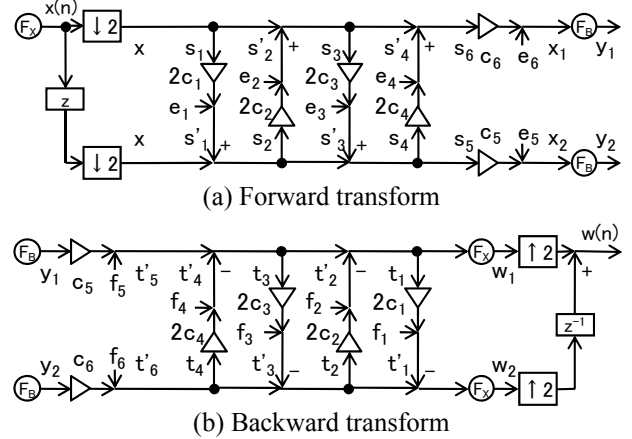


Fig. 2 Equivalent circuits of the DWT for a DC input.

C. Nullification of Accumulated Errors

In Fig.2(a), the unified errors in Eq.(15) are propagated and accumulated inside the circuit. When the accumulated errors are nullified by the rounding at output of the forward transform, Eq.(9) is satisfied. In the figure, $\mathbf{Y}_{12} = [y_1 \ y_2]^T$ and $\mathbf{W}_{12} = [w_1 \ w_2]^T$ are described as

$$\begin{cases} \mathbf{Y}_{12} = R_{F_B} [(\mathbf{H}_{e1}\mathbf{E}_1 + \mathbf{H}_{e2}\mathbf{E}_2) + \mathbf{KH}_{4321}\mathbf{X}] \\ \mathbf{W}_{12} = R_{F_X} [(\mathbf{H}_{e3}\mathbf{E}_3 + \mathbf{H}_{e4}\mathbf{E}_4) + (\mathbf{KH}_{4321})^{-1}\mathbf{Y}_{12}] \end{cases} \quad (16)$$

where the errors are

$$\begin{cases} \mathbf{E}_1 = [e_6 & e_4 & e_2]^T \\ \mathbf{E}_2 = [e_5 & e_3 & e_1]^T \end{cases} \begin{cases} \mathbf{E}_3 = [f_2 & f_4 & f_5]^T \\ \mathbf{E}_4 = [f_1 & f_3 & f_6]^T \end{cases}$$

and transfer function to the errors are

$$\begin{cases} \mathbf{H}_{e1} = [\mathbf{I}_U & \mathbf{K}\mathbf{I}_U & \mathbf{K}\mathbf{H}_{43}\mathbf{I}_U] \\ \mathbf{H}_{e2} = [\mathbf{I}_L & \mathbf{K}\mathbf{H}_4\mathbf{I}_L & \mathbf{K}\mathbf{H}_{432}\mathbf{I}_L] \end{cases} \begin{cases} \mathbf{H}_{e3} = -[\mathbf{H}_1^{-1}\mathbf{I}_U & \mathbf{H}_{123}^{-1}\mathbf{I}_U & \mathbf{H}_{1234}^{-1}\mathbf{I}_U] \\ \mathbf{H}_{e4} = -[\mathbf{I}_L & \mathbf{H}_{12}^{-1}\mathbf{I}_L & \mathbf{H}_{1234}^{-1}\mathbf{I}_L] \end{cases}$$

In the equations above, we used the notations:

$$\mathbf{H}_{i \in \{1,3\}} = \begin{bmatrix} 1 & 0 \\ 2c_i & 1 \end{bmatrix}, \quad \mathbf{H}_{j \in \{2,4\}} = \begin{bmatrix} 1 & 2c_j \\ 0 & 1 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} c_6 & 0 \\ 0 & c_5 \end{bmatrix},$$

$$\mathbf{H}_{43 \dots i} = \mathbf{H}_4 \mathbf{H}_3 \dots \mathbf{H}_i, \quad \mathbf{H}_{12 \dots i}^{-1} = \mathbf{H}_1^{-1} \mathbf{H}_2^{-1} \dots \mathbf{H}_i^{-1},$$

$$\mathbf{I}_U = [1 \ 0]^T, \quad \mathbf{I}_L = [0 \ 1]^T, \quad \mathbf{I}_{UL} = \mathbf{I}_U + \mathbf{I}_L.$$

Subtracting the ideal output from the output in Eq.(16), we have the accumulated errors as

$$\begin{bmatrix} \mathbf{E}_{y12} \\ \mathbf{E}_{w12} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{12} \\ \mathbf{W}_{12} \end{bmatrix} - \begin{bmatrix} \mathbf{I}_U \\ \mathbf{I}_{UL} \end{bmatrix} x. \quad (17)$$

Finally, using the property in Eq.(8), we have

$$\begin{bmatrix} \mathbf{E}_{y12} \\ \mathbf{E}_{w12} \end{bmatrix} = \begin{bmatrix} R_{F_B} [(\mathbf{H}_{e1}\mathbf{E}_1 + \mathbf{H}_{e2}\mathbf{E}_2)] \\ R_{F_X} [(\mathbf{H}_{e3}\mathbf{E}_3 + \mathbf{H}_{e4}\mathbf{E}_4)] \end{bmatrix}. \quad (18)$$

Applying Eq.(7) with $p=0$, it becomes clear that when the conditions:

$$\begin{cases} \|\mathbf{H}_{e1}\mathbf{E}_1 + \mathbf{H}_{e2}\mathbf{E}_2\| \leq \mathbf{I}_{UL} 2^{-1-F_B} \\ \|\mathbf{H}_{e3}\mathbf{E}_3 + \mathbf{H}_{e4}\mathbf{E}_4\| \leq \mathbf{I}_{UL} 2^{-1-F_X} \end{cases} \quad (19)$$

are satisfied, the accumulated errors are nullified by the rounding operations at output of the transforms.

D. Critical Condition on Word Length

Based on discussions above, we derive the condition on word length of coefficients and signals such that the condition in Eq.(19) is satisfied. Since the unified errors in \mathbf{E}_1 , \mathbf{E}_2 , \mathbf{E}_3 and \mathbf{E}_4 have the maximum in Eq.(15) described with the parameters p_i and q_i , the DC lossless condition is also described with the parameters by substituting

$$\begin{cases} \mathbf{E}_1 = \left([p_6 \ p_4 \ p_2]^T + \mathbf{I}_3 \cdot 2^{-1} \right) 2^{-F_S} \\ \mathbf{E}_2 = \left([p_5 \ p_3 \ p_1]^T + \mathbf{I}_3 \cdot 2^{-1} \right) 2^{-F_S} \\ \mathbf{E}_3 = \left([q_2 \ q_4 \ q_5]^T + \mathbf{I}_3 \cdot 2^{-1} \right) 2^{-F_S} \\ \mathbf{E}_4 = \left([q_1 \ q_3 \ q_6]^T + \mathbf{I}_3 \cdot 2^{-1} \right) 2^{-F_S} \end{cases} \quad (20)$$

into Eq.(19) where $\mathbf{I}_3 = [1 \ 1 \ 1]^T$. This is the condition we derived based on the model in III.A.

In IV, we investigate the minimum F_{C_i} [bit] of a coefficient c_i , $i \in I$ under the DC lossless condition for a given DC value x and F_S [bit] of signals inside the DWT.

E. Sufficient Condition on Word Length

In case of all the parameter in Eq.(20) are given as $p_i = q_i = p$ and $F_{C_i} = F_C$ for $\forall i \in I$, the condition in Eq.(19) becomes

$$\begin{cases} \left(\|\mathbf{H}_{e1}\|_{L^1} + \|\mathbf{H}_{e2}\|_{L^1} \right) \cdot 2^{-F_S} (p + 2^{-1}) \leq \mathbf{I}_{UL} 2^{-1-F_B} \\ \left(\|\mathbf{H}_{e3}\|_{L^1} + \|\mathbf{H}_{e4}\|_{L^1} \right) \cdot 2^{-F_S} (p + 2^{-1}) \leq \mathbf{I}_{UL} 2^{-1-F_X} \end{cases} \quad (21)$$

where $\|\mathbf{H}\|_{L^1}$ denotes a column vector whose component is a sum of absolute value of all components in each row. Substituting coefficients of the DWT into Eq.(21), we have

$$p \leq 2^{-1+F_S-G_E} - 2^{-1}, \quad G_E = 2.66 \text{ [bit]} \quad (22)$$

for $F_X = F_B = 0$. As a result, the DC lossless condition on the word length is given as

$$-\log_2(2^{-\Delta W_C} + 2^{-\Delta W_S}) \geq G_E \quad (23)$$

where

$$[\Delta W_C \ \Delta W_S] = [F_C - I_S \ F_S]$$

and G_E is the lower bound. This means a sufficient condition for the DC lossless. Since it is too strict, the word length under this condition is redundant. Unlike this sufficient condition, our critical condition given as Eq.(19) under Eq.(20) determines the word length minimum and necessary for the DC lossless.

IV. SIMULATION RESULTS

A. Word Length under the Conditions

In Fig.3, the sufficient condition in III.E is plotted as 'solid' line for any 8 bit integer x ($W_X = I_X = 8$, $F_X = 0$). The optimum word length under this condition is $(F_S, F_C) = (4, 12)$ as indicated as 'Existing'. It guarantees the DC lossless, however the condition is too strict. Therefore the word length is redundant and there is room for further reduction.

In Fig.3, a cross 'x' indicates a pair (F_S, F_C) which satisfies the critical condition in III.D. The minimum of F_C for each F_S is indicated as a 'broken' line. It is clear that the word length derived by our critical condition is shorter than that by the sufficient condition based on a conventional approach.

Table I summarizes the minimum word length under the conditions. It is observed that the word length can be reduced by limiting input DC signals to a specific value. Unlike the conventional analysis, our analysis gives the minimum word length shorter than that determined by the sufficient condition for each of input DC values.

Table I The minimum word length for a specific value for white balancing of a video system.

condition	sufficient	critical	
		$x_{in} = 16$ (black)	$x_{in} = 235$ (white)
F_S (signal)	4	2	3
F_C (coefficient)	12	12	9

$$x_{in} = x - 2^{I_X - 1}$$

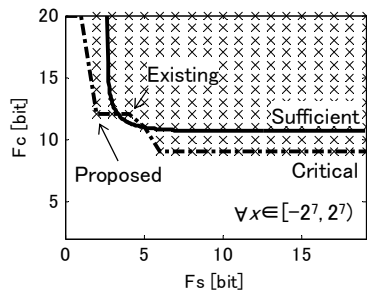


Fig.3 Word length under the conditions. "x" indicates (F_s, F_c) such that the DWT becomes DC lossless.

Table II Tolerance parameters for $x_m=16$ and $F_s=2$.

F_c	p_1	p_2	p_3	p_4	p_5	p_6	y_1-x	y_2
9	0	0	0	0	0	1	0	0
8	0	-3	0	0	0	0	-1	-1
7	0	-3	0	0	0	0	-1	-1
6	7	12	-8	0	0	2	6	6
5	7	-18	10	0	1	0	-7	-5
4	-21	-18	9	0	-2	1	-10	-12
3	35	107	7	25	9	-28	63	70

Table III The minimum word length of coefficients of the 9-7 DWT for specific values.

input values		signals		coefficients					
		F_s	F_{C1}	F_{C2}	F_{C3}	F_{C4}	F_{C5}	F_{C6}	ave.
for- ward	$x_m=16$ (B)	2	7	9	7	4	6	4	6.17
	$x_m=235$ (W)	3	9	7	9	9	1	4	6.50
back- ward	$x_m=16$ (B)	2	9	9	7	0	8	0	5.50
	$x_m=235$ (W)	3	9	9	7	0	8	0	5.50

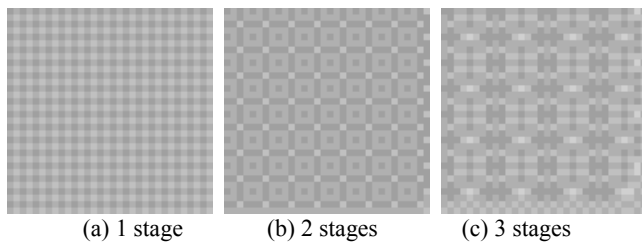


Fig.4 Example of reconstructed images for 128^2 pixel DC input image with $x=10$. Intensity is multiplied by 16.

B. Optimum Word Length Assignment

Since we described tolerance for the unified errors as parameters p_i and q_i in Eq.(15), it becomes possible to simultaneously control both of word length of signals and that of coefficients.

Table II summarizes these parameters for an input value 16 and the word length of signals at $F_s=2$ [bit] as an example. It indicates that $[p_1 p_2 \dots p_6]$ are $[0 0 0 0 0 1]$ for $F_c=9$ [bit]. In this case, all the coefficients c_i in the forward transform have the same length. Note that the parameter p_1 is the same for $F_c=9, 8$ and 7 [bit]. It means that word length of the coefficient c_1 can be reduced from 9 to 7 [bit] without any influence to the errors. As a result, $[F_{C1} F_{C2} \dots F_{C6}]$ can be reduced from $[9 9 9 9 9 9]$ to $[7 9 7 4 6 4]$.

Table III summarizes results of this optimum word length assignment for the DWT. Comparing to table I, it is observed that word length of coefficients is reduced from 9.00 [bit] to 6.17 [bit] on average for an input value $x_m=16$.

Fig.4 illustrates image signals reconstructed by the DWT which does not satisfy the DC lossless condition. It demonstrates the checker board artifact for reference. A DWT with the coefficients under the condition described in this report does not have such artifacts.

V. CONCLUSIONS

Utilizing the nullification of the accumulated errors, we theoretically derived a condition on word length of signals and coefficients such that the 9-7 DWT of JPEG 2000 becomes lossless for a DC input signal. It was confirmed that the minimum word length derived by our 'critical' condition was shorter than that determined by a conventional 'sufficient' condition. Analysis in this report contributes to build a low complexity DC lossless DWT.

REFERENCES

- [1] ISO/IEC FCD15444-1, "JPEG2000 image coding system," March 2000.
- [2] H. Kiya, M. Yae, M. Iwahashi, "Linear phase two channel filter bank allowing perfect reconstruction", IEEE Proc. international symposium on circuits and systems (ISCAS), no.2, pp.951-954, May 1992.
- [3] W. Sweldens, "The lifting scheme: A custom-design construction of biorthogonal wavelets," Technical Report 1994:7, industrial mathematics initiative, department of mathematics, university of South Carolina, 1994.
- [4] A. M. Reza, Lian Zhu, "Analysis of error in the fixed-point implementation of two-dimensional discrete wavelet transforms," IEEE Trans. circuits and systems, fundamental theory and applications, vol.52, issue 3, pp.641-655, March 2005.
- [5] H. Kiya, M. Iwahashi, O. Watanabe, "A new class of lifting wavelet transform for guaranteeing losslessness of specific signals," IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp.3273-3276, March 2008.
- [6] Y. Harada, S. Muramatsu, H. Kiya, "Two channel QMF bank without checker board effect and its lattice structure," IEICE Trans. on fundamentals, vol.J80-A, no.11, pp.1857-1867, Nov. 1997.
- [7] Y. Tanaka, M. Ikehara, "First order linear phase filter banks with regularity constrains for efficient image coding," IEICE Trans. fundamentals, vol. J91-A, no.2, pp.192-201, Feb. 2008.
- [8] Wei Dai, T. D. Tran, "Regularity-constrained pre- and post- filtering for block DCT-based systems," IEEE Trans. signal processing, vol.51, Issue 10, pp.2568-2581, Oct. 2003.
- [9] K. Hirakawa, T. W. Parks, "Chromatic adaptation and white balancing problem," IEEE Proc. international conference on image processing (ICIP), vol. III, pp.984-987, Nov. 2005.
- [10] M. Iwahashi, H. Kiya, "Word length condition for DC Lossless DWT," Asia pacific signal and information processing association (APSIPA) annual summit and conference, no.TA-P2-6, pp.469-472, Oct. 2009.
- [11] M. Iwahashi, H. Kiya, "Finite word length error analysis based on basic formula of rounding operation", the international symposium on intelligent signal processing and communication systems (ISPACS), no.86, pp.49-52, Dec. 2008.
- [12] The society of motion picture and television engineers, "Standard for television, 1920x1080 image sample structure, digital representation and digital timing reference sequences for multiple picture rates", SMPTE 274 M-2005, Feb.2005.