CCITT SG-XV
Working Party XV/1
Specialists Group on Coding for Visual Telephony

Document #222
June 1987

Source: NTT, KDD, NEC and FUJITSU
Title : PRECISION OF DCT CALCULATION

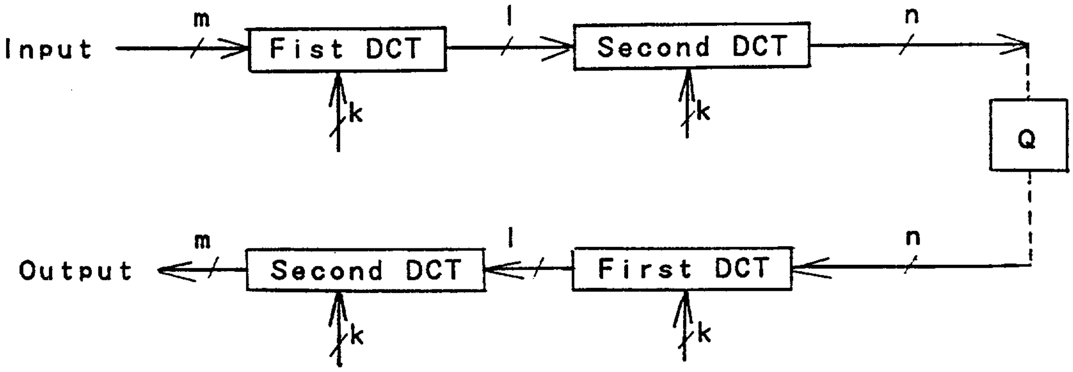
---

1. Introduction

According to the current Flexible Hardware specification, accuracy of Inverse DCT calculation is strictly specified to ensure compatibility between independently manufactured programmable codecs [1]. It is specified that Inverse DCT calculation is carried out with 16-bit accuracy for matrix elements and 16-bit accuracy for the first one-dimensional DCT output to be supplied to the second one in the cascaded DCT [2]. To facilitate DCT implementation, appropriate accuracy should be investigated [3].

This contribution shows experimental results when these accuracies (number of bits) are changed in the calculation using actual video sequences.

2. Accuracy of Forward and Inverse DCT ---- Definition



In the current Flexible Hardware specification, the values for m, k, l, and n are 9, 16, 16, and 12 bits, respectively. It is noted that l bits are obtained by truncation from the first DCT output expressed with (m+k-1) or (n+k-1) bits by truncation.

3. Experiments Using Edge Pattern

(1) Truncation or Round-off ? (Table 1)

In this experiment when no quantizer is used, approximation is made at a point between the first and the second DCT for both Forward and Inverse DCT. The result is shown in Table 1. For a combination of m=9, k=12, l=8, and n=12, round-off approximation gives a better result in that distortion by round-off is apparently much less biased than by truncation. No doubt that the difference is small when the accuracy of the parameters is high.

(2) Quantizer and DCT Accuracy (Table 2)

For a quantizer with (4, 4) i.e. a threshold and a step size both equal to 4 in RM4 but in open-loop operation, difference is hardly found even if the accuracy for (k, l) is changed from (16, 16) to (16, 10). For a coarser quantizer with (16, 16), difference is hardly seen for the same parameter change. However, when k and l are taken to be 12 and 8, respectively, distortion increases slightly. Round-off is used for approximation.

It seems that a combination of (k, l) equal to (16, 16) is more than enough as far as this experiment is concerned.

4. Experiment Using 'Split Screen' (Fig.1)

Maximum accuracy here is given by a calculation which fully meets the Flexible Hardware specification. SNR values are shown as a function of (k, l) values in Fig. 1.

Picture quality difference is hardly found for (k, l) values of (12, 12), (10, 12), (10, 10), and (8, 10) when the two quantizers above i.e. (threshold, step size)= (4, 4) and (16, 16) are used. Round-off is employed in approximating the output value from the first DCT.

Picture quality degradation is apparent when the accuracy for this output value is reduced to 8 bits. For reduced accuracy of this output, round-off is in particular effective [4].

The 16-bit accuracy for k and l seems more than enough.

Conclusion

An appropriate accuracy for DCT calculation seems to be 12 bits for both k and l, though further close investigation is needed. As an approximation in DCT calculation, round-off is recommended. (VTR demonstration follows.)

Reference

[1] Doc.#182 (Nov. 1986)
[2] Doc.#142 (Sept. 1986)
[3] Doc.#121 (June 1986)
[4] Section 4.2/Doc.#181R (Nov. 1986)

(ORIGINAL PICTURE)

0	0	0	0	0	0	128	128
0	0	0	0	0	128	128	128
0	0	0	0	0	128	128	128
0	0	0	0	128	128	128	128
0	0	0	0	128	128	128	128
0	0	0	128	128	128	128	128
0	0	0	128	128	128	128	128
0	0	128	128	128	128	128	128

Forward	m	k	l	n
	9	12	8	12
	9	12	8	12
Inverse	m	k	l	n

First DCT $\xrightarrow{\text{Round-off}}$ Second DCT

(ORIGINAL PICTURE)-(DECODED PICTURE)

11	-4	9	10	18	8	1	-2
11	1	2	-1	-8	-8	1	1
12	-6	4	1	0	7	3	2
10	-4	0	8	1	-2	2	-1
9	-5	7	4	-3	4	1	-2
12	-1	-1	4	3	2	7	3
11	-2	11	9	2	1	-2	0
12	0	4	-4	4	2	5	-1

First DCT $\xrightarrow{\text{Truncation}}$ Second DCT

(ORIGINAL PICTURE)-(DECODED PICTURE)

5	3	0	2	-3	-2	2	0
2	0	0	-1	0	-3	-1	3
-2	-2	2	2	4	-2	-3	0
-3	0	-6	0	1	-3	5	-2
1	-1	0	1	2	3	4	3
1	3	2	-3	-1	-1	2	3
4	1	-1	0	1	-4	0	6
0	2	-1	5	0	-3	1	-5

Table 1 Approximation

Quantizer
(Threshold, Step Size) = (4.0, 4.0)

Forward	m	k	l	n
	9	16	16	12
	9	16	16	12
Inverse	m	k	l	n

(OGIGINAL PICTURE)-(DECODED PICTURE)

-2	0	1	0	0	1	0	0
0	1	-1	1	0	1	-1	0
1	1	0	0	1	0	-1	-1
0	1	-1	0	-2	1	0	-1
1	-1	-1	2	-1	0	-2	0
1	1	-1	-2	-1	0	-1	-1
-1	1	-2	0	-2	0	-2	-1
0	-1	-1	-1	0	-1	-1	1

Forward	m	k	l	n
	9	12	8	12
	9	12	8	12
Inverse	m	k	l	n

(ORIGINAL PICTURE)-(DECODED PICTURE)

1	-1	-4	0	1	1	-1	-1
-1	0	-2	0	-1	0	-3	0
1	0	1	2	1	-2	-3	1
0	0	0	2	-1	1	1	0
-1	-1	-1	0	-2	-1	-1	-1
-1	3	2	-1	-2	-1	0	-1
1	3	0	1	0	2	0	1
0	1	-1	-2	0	3	0	-2

Table 2 (a) Accuracy

Quantizer
(Threshold, Step Size) = (16.0, 16.0)

Forward	m	k	l	n
	9	16	16	12
	9	16	16	12
Inverse	m	k	l	n

(ORIGINAL PICTURE)-(DECODED PICTURE)

-6	-2	-4	2	-9	5	2	-6
7	0	4	6	0	-3	3	0
-6	-6	1	-2	-5	1	2	-8
0	-7	-7	-4	0	-1	-1	2
-4	-1	-1	-2	2	4	5	-2
6	-4	-3	3	0	-3	4	4
-2	-5	1	-2	-9	-6	-2	-9
4	-4	-7	7	-4	2	0	4

Forward	m	k	l	n
	9	12	10	12
	9	12	10	12
Inverse	m	k	l	n

(ORIGINAL PICTURE)-(DECODED PICTURE)

-10	1	-4	3	-8	5	2	-5
8	-1	3	3	-4	-6	-1	-3
-3	-3	5	3	-2	0	4	-6
-2	-9	-8	-2	-1	-5	-4	0
-2	2	3	-2	-1	6	7	0
3	-6	-3	0	-5	-7	1	1
1	-1	4	2	-5	-5	-1	-10
3	-4	-7	6	-5	2	-3	7

Table 2 (b) Accuracy

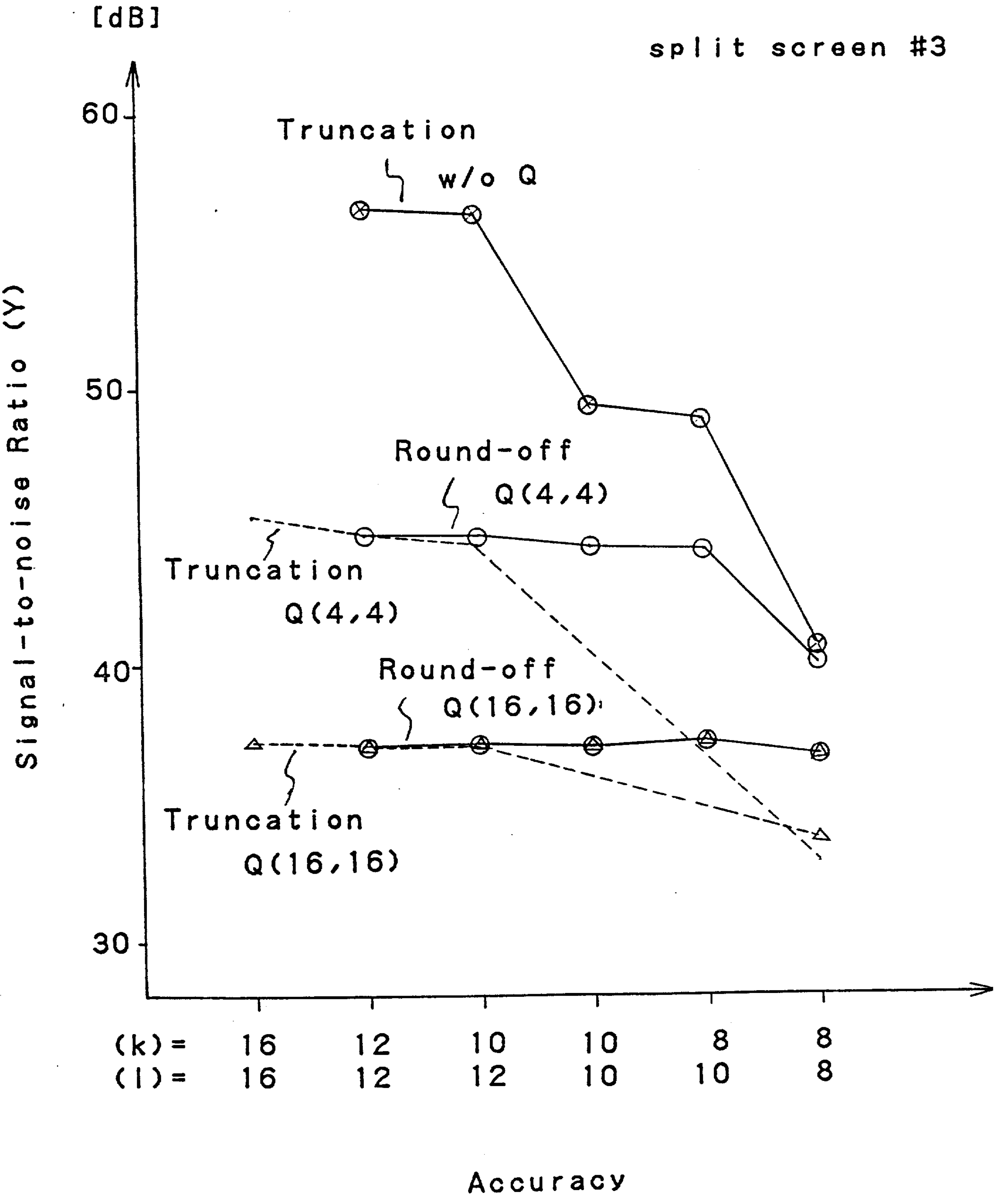


Figure 1 Accuracy and Distortion