

M. Zhou ¹ and J. De Lameillieure

Department of Image Processing
Heinrich-Hertz-Institut für Nachrichtentechnik Berlin GmbH
Einsteinufer 37, D - 10587 Berlin, Germany

Abstract

The precision of the IDCT output before clipping in a MPEG hardware decoder should be sufficient to cover the maximum output range of the IDCT transform. This paper proves theoretically that the IDCT output range of MPEG video coding is $[-1805, 1805]$ with the quantization of MPEG Test Model 5 [2], examples approaching the upper and lower bound of the IDCT output have verified the theoretical analysis. The experimental results with the realistic quantization matrices and reasonable input block vectors are also presented in this paper. The theoretical analysis and the simulation results show that at least 11 bit and maximal 12 bit should be used to represent the IDCT output of MPEG video coding with TM5 quantizer before clipping in order to avoid overflow.

Key words: MPEG, IDCT output range, Quantization factor

1 Introduction

In the common applications of digital image coding like digital TV, videotelephony and recently multimedia applications, the pel values have a precision of 8 bit. In the DCT-coding of 8-bit image material, e.g. in MPEG coding, the input of the DCT transform should have a precision of 8 bit in "intra"-coding and 9 bit in motion compensated predictive coding².

However, after quantization and inverse quantization in MPEG coding and decoding, the range of the output pel values can be larger, much larger, as we will show later in this paper. As a consequence, the precision of the IDCT output and the internal IDCT precision should be larger than the necessary precision at the DCT input.

In MPEG coding, the IDCT output is clipped to a range of $[-255, 255]$. Sometimes, the inverse discrete cosine transform and the clipping is considered as one process, called "IDCT". In this paper, we always make a distinction between the transformation itself, in the following called "IDCT", and the clipping. A block scheme with the relevant parts of the MPEG coder and decoder is given in Fig. 1.

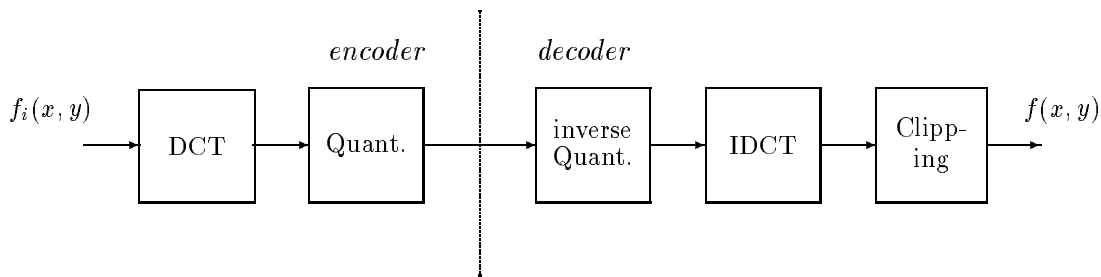


Figure 1: Block scheme of the signal path through DCT, quantization, inverse quantization, IDCT and clipping in an MPEG encoder and decoder; The input $f_i(x, y)$ are 2D pel values in intra-coded blocks and 2D prediction errors in predictive blocks; The output $f(x, y)$ corresponds to the $f(x, y)$ of the MPEG video IS [1].

There is a penalty in decoder cost for the extra necessary IDCT precision. As it is the goal of the MPEG standardisation group to standardise a decoder at a reasonable cost, it has been discussed extensively in the MPEG group what should be the sufficient precision of a compliant decoder [6].

¹Corresponding author. Phone: +49-30-31002 616; fax : +49-30-392 72 00; e-mail: zhou@hhi.de

²"Predictive" coding here means as well pure forward predictive coding as bidirectional coding.

In the MPEG group, some bitstreams have been generated that produce enormous IDCT outputs. However, these bitstreams have not been generated by encoders that compress an input video signal. Indeed, MPEG, especially MPEG-2, allows to represent extremely large DCT coefficients in the bitstream, large DCT coefficients that could not be the result of DCT transformation of a block with pel values in the range $[-255, 255]$. According to the discussions in the MPEG compliance group, these bitstreams belong to the so-called “twilight” zone.

This paper is an investigation of the maximal output range of the IDCT. It contains first a theoretical derivation of the upper bound of the maximal range and then examples that illustrate the relevance of the theoretical bound.

The aim of this paper is to be a basis for the decision on the necessary IDCT precision before clipping in MPEG video coding, decisions that should be reflected in the video[1] and the compliance standard[3]

This paper is organized as follows: In section 2 the $N \times N$ two-dimensional inverse discrete cosine transform IDCT is briefly reviewed and two lemmas about IDCT are proved, in section 3 the general form of IDCT output range of MPEG video coding with a constant maximum quantizer factor is derived, in section 4 the TM5 quantizer [2] is intensively analyzed and then the IDCT output range of MPEG video coding with the TM5 quantizer is proved, section 5 presents some examples to verify the theoretical analysis and to demonstrate what the IDCT output can be with the realistic quantization matrices and the reasonable input block vectors, followed by the conclusion in section 6.

2 Inverse Discrete Cosine Transform (IDCT)

The Discrete Cosine Transform (DCT) is the most popular technique for image transform coding. The $N \times N$ two dimensional Inverse Discrete Cosine Transform IDCT is defined as

$$f(x, y) = \frac{2}{N} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} C(u)C(v)F(u, v) \cos \frac{(2x+1)u\pi}{2N} \cos \frac{(2y+1)v\pi}{2N}$$

where $u, v, x, y = 0, 1, 2, \dots, N-1$. The variables x, y are spatial coordinates in the sample domain, while u, v are coordinates in the transform domain. The function $C(\cdot)$ is defined as

$$C(u), C(v) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } u, v=0 \\ 1 & \text{otherwise} \end{cases}$$

By a rearrangement of

$$\begin{cases} G_{u \cdot N + v} & = F(u, v) \\ g_{x \cdot N + y} & = f(x, y) \\ T_{x \cdot N + y, u \cdot N + v} & = \frac{2}{N} C(u)C(v) \cos \frac{(2x+1)u\pi}{2N} \cos \frac{(2y+1)v\pi}{2N} \end{cases}$$

the two-dimensional IDCT can be written as an one-dimensional linear transform

$$\vec{g} = T\vec{G} \tag{1}$$

where $\vec{g} = (g_0, g_1, g_2, \dots, g_{N^2-1})^t$ and $\vec{G} = (G_0, G_1, G_2, \dots, G_{N^2-1})^t$ denote the input block vector and the DCT-coefficient vector after the rearrangement, respectively. Here t (and hereafter) denotes transpose. The matrix

$$T = \begin{pmatrix} T_{0,0} & T_{0,1} & \cdots & T_{0,N^2-1} \\ T_{1,0} & T_{1,1} & \cdots & T_{1,N^2-1} \\ \vdots & \vdots & \vdots & \vdots \\ T_{N^2-1,0} & T_{N^2-1,1} & \cdots & T_{N^2-1,N^2-1} \end{pmatrix}$$

is the $N^2 \times N^2$ -dimensional transform matrix, where $T_{i,k} \neq 0, \forall i, k = 0, 1, 2, \dots, N^2 - 1$.

It is well known that the transform matrix T of the $N \times N$ two dimensional IDCT is orthonormal [7], therefore it satisfies

$$\sum_{k=0}^{N^2-1} G_k^2 = \sum_{i=0}^{N^2-1} g_i^2 \quad (2)$$

and

$$\sum_{k=0}^{N^2-1} T_{i,k} \cdot T_{j,k} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{otherwise} \end{cases} \quad \forall i, j = 0, 1, 2, \dots, N^2 - 1 \quad (3)$$

As a preparation of the theoretical deviation of the following sections, a lemma is proved here.

Lemma 1 For a $N \times N$ input block vector $\vec{g} = (g_0, g_1, g_2 \dots g_{N^2-1})^t$ with its elements in the interval $[a, b]$ ($b > a$), after $N \times N$ DCT its DCT-coefficient vector $\vec{G} = (G_0, G_1, G_2 \dots G_{N^2-1})^t$ satisfies

$$\sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k| \leq \frac{|a+b|}{2} + N \cdot \frac{b-a}{2} \quad \forall i = 0, 1, 2, \dots, N^2 - 1$$

where $|\cdot|$ is the absolute operator.

Proof (Lemma 1): The DCT-coefficient vector \vec{G} can be expressed as the sum of two vectors, that is

$$\vec{G} = \vec{D} + \vec{R}$$

where $\vec{D} = (N \cdot \frac{a+b}{2}, 0, 0 \dots 0)^t$, $\vec{R} = (G_0 - N \cdot \frac{a+b}{2}, G_1, G_2 \dots G_{N^2-1})^t$.

Note that only the first element of \vec{D} has non-zero value, the sum $\sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k|$ can be rewritten as

$$\begin{aligned} \sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k| &= |T_{i,0}| \cdot (|G_0| - |G_0 - N \cdot \frac{a+b}{2}|) \\ &\quad + \sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |R_k| \end{aligned} \quad (4)$$

Because $|G_0| - |G_0 - N \cdot \frac{a+b}{2}| \leq N \cdot \frac{|a+b|}{2}$ and $T_{i,0} = \frac{1}{N}$, $\forall i = 0, 1, 2, \dots, N^2 - 1$, we have

$$\sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k| \leq \frac{|a+b|}{2} + \sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |R_k|$$

Considering that the inner product of two vectors is always not larger than the amplitude product of these two vectors, we get

$$\sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k| \leq \frac{|a+b|}{2} + \sqrt{\sum_{k=0}^{N^2-1} T_{i,k}^2} \cdot \sqrt{\sum_{k=0}^{N^2-1} R_k^2}$$

Seeing that \vec{R} is actually the DCT-coefficient vector of the input block vector $\vec{r} = (g_0 - \frac{a+b}{2}, g_1 - \frac{a+b}{2}, g_2 - \frac{a+b}{2} \dots g_{N^2-1} - \frac{a+b}{2})^t$, we deduce the following formula by making use of the orthonormality of the IDCT given in (2) and (3)

$$\sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k| \leq \frac{|a+b|}{2} + \sqrt{\sum_{k=0}^{N^2-1} (g_k - \frac{a+b}{2})^2} \quad (5)$$

As the elements of the input vector \vec{g} are in the interval $[a, b]$, the left side of formula (5) has the maximum value $\frac{a+b}{2} + N \cdot \frac{b-a}{2}$ when $g_k = a$ or b , for $k = 0, 1, 2, \dots, N^2 - 1$. Hence, we have proved

$$\sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k| \leq \frac{|a+b|}{2} + N \cdot \frac{b-a}{2} \quad \forall i = 0, 1, 2, \dots, N^2 - 1$$

Now we will investigate the form of the DCT-coefficient vector \vec{G} when the sum $\sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k|$ reaches its maximum value $\frac{|a+b|}{2} + N \cdot \frac{b-a}{2}$.

The conditions that let the sum $\sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k|$ have the maximum value can be acquired from (4) and (5), they are

$$\begin{cases} |G_0| - |G_0 - N \cdot \frac{a+b}{2}| = N \cdot \frac{|a+b|}{2} \\ g_k = a \text{ or } b \\ |R_k| = N \cdot \frac{b-a}{2} \cdot |T_{i,k}| \end{cases} \quad \begin{matrix} \text{for } k = 0, 1, 2, \dots, N^2 - 1 \\ \\ \text{for } k = 0, 1, 2, \dots, N^2 - 1 \end{matrix} \quad (6)$$

Because $R_0 = G_0 - N \cdot \frac{a+b}{2}$ and $T_{i,0} = \frac{1}{N}$ it can be derived from (6) that \vec{G} must have the form

$$\begin{cases} G_0 = \begin{cases} N \cdot \frac{a+b}{2} + \frac{b-a}{2} & \text{if } a+b > 0 \\ N \cdot \frac{a+b}{2} - \frac{b-a}{2} & \text{if } a+b < 0 \\ \pm \frac{b-a}{2} & \text{if } a+b = 0 \end{cases} \\ |G_k| = N \cdot \frac{b-a}{2} \cdot |T_{i,k}| \quad \text{for } k \neq 0 \end{cases}$$

According to the deviation above, we have

Lemma 2 For a $N \times N$ input block vector $\vec{g} = (g_0, g_1, g_2 \dots g_{N^2-1})^t$ with its elements in the interval $[a, b]$ ($b > a$), if its DCT-coefficient vector $\vec{G} = (G_0, G_1, G_2 \dots G_{N^2-1})^t$ satisfies

$$\sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k| = \frac{|a+b|}{2} + N \cdot \frac{b-a}{2}$$

the vector \vec{G} has the form

$$\begin{cases} G_0 = \begin{cases} N \cdot \frac{a+b}{2} + \frac{b-a}{2} & \text{if } a+b > 0 \\ N \cdot \frac{a+b}{2} - \frac{b-a}{2} & \text{if } a+b < 0 \\ \pm \frac{b-a}{2} & \text{if } a+b = 0 \end{cases} \\ |G_k| = N \cdot \frac{b-a}{2} \cdot |T_{i,k}| \quad \text{for } k \neq 0 \end{cases}$$

3 IDCT output range of MPEG video coding with a constant maximum quantization factor

In order to reduce the bit-rate of an image sequence the DCT-coefficients of a block have to be quantized. At the decoder the received DCT coefficients are inversely quantized and transformed by the IDCT. The quantization of DCT-coefficients degrades the quality of the decoded image sequence. On top of that, it is responsible for the possible extreme IDCT outputs.

The problem of obtaining the IDCT output range of MPEG video coding is equivalent to looking for the upper bound and lower bound of the IDCT output. For this purpose we first investigate a general case. Let \vec{g} be an input block vector with its elements in the interval $[a, b]$ ($b > a$) and \vec{G} be its corresponding DCT-coefficient vector. $\tilde{\vec{G}} = (\tilde{G}_0, \tilde{G}_1, \tilde{G}_2 \dots \tilde{G}_{N^2-1})^t$ denotes the reconstructed DCT-coefficient vector after quantization and inverse quantization of vector \vec{G} , and $\tilde{\vec{g}} = (\tilde{g}_0, \tilde{g}_1, \tilde{g}_2 \dots \tilde{g}_{N^2-1})^t$ represents the reconstructed block vector (e.g. IDCT output) of \vec{g} obtained by making the $N \times N$ two dimensional IDCT of $\tilde{\vec{G}}$. Here, the upper bound and lower bound of the IDCT output are the maximum value and minimum value the elements of $\tilde{\vec{g}}$ can reach, respectively.

The process of quantization and inverse quantization of a DCT-coefficient G_k can be simply modeled as

$$\hat{G}_k = \eta_k \cdot G_k \quad \text{for } k = 0, 1, 2, \dots, N^2 - 1 \quad (7)$$

where η_k is the so called "quantization factor" of DCT-coefficient G_k . If we assume that the used quantizers are the reasonable ones which never quantize and reconstruct a DCT-coefficient as a coefficient of opposite sign, then $\eta_k \geq 0$. This is the case in the TM5 and SM3 quantization [2][5].

Let \hat{g}_i be an arbitrary element of vector $\tilde{\mathbf{g}}$ with $i = 0, 1, 2, \dots, N^2 - 1$, according to (1) and (7) we have

$$\hat{g}_i = \sum_{k=0}^{N^2-1} T_{i,k} \cdot G_k \cdot \eta_k \quad (8)$$

In order to let \hat{g}_i in (8) have the maximum value we select the quantization factors as

$$\eta_k = \begin{cases} \eta_{max} & \text{for } T_{i,k} \cdot G_k > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } k = 0, 1, 2, \dots, N^2 - 1 \quad (9)$$

where η_{max} is the maximum quantization factor. The maximum quantization factor depends on the used quantizer and the inverse quantization. Here we assume that the maximum quantization factor is a constant.

For the convenience of derivation, we rewrite (9) as

$$\eta_k = \begin{cases} (1 + \frac{T_{i,k} \cdot G_k}{|T_{i,k} \cdot G_k|}) \cdot \frac{\eta_{max}}{2} & \text{for } G_k \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } k = 0, 1, 2, \dots, N^2 - 1 \quad (10)$$

Substituting η_k of (10) into (8) we have

$$\begin{aligned} \hat{g}_i &\leq (\sum_{k=0}^{N^2-1} T_{i,k} \cdot G_k + \sum_{k=0}^{N^2-1} |T_{i,k} \cdot G_k|) \cdot \frac{\eta_{max}}{2} \\ &= (g_i + \sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k|) \cdot \frac{\eta_{max}}{2} \end{aligned} \quad (11)$$

Using Lemma 1 and selecting $g_i = b$ we obtain the upper bound of the $N \times N$ IDCT output from (11)

$$\hat{g}_i^{upper} \leq (b + \frac{|a+b|}{2} + N \cdot \frac{b-a}{2}) \cdot \frac{\eta_{max}}{2}$$

Similarly, by setting

$$\eta_k = \begin{cases} (1 - \frac{T_{i,k} \cdot G_k}{|T_{i,k} \cdot G_k|}) \cdot \frac{\eta_{max}}{2} & \text{for } G_k \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

for $k = 0, 1, 2, \dots, N^2 - 1$

we can easily get

$$\begin{aligned} \hat{g}_i &\geq (\sum_{k=0}^{N^2-1} T_{i,k} \cdot G_k - \sum_{k=0}^{N^2-1} |T_{i,k} \cdot G_k|) \cdot \frac{\eta_{max}}{2} \\ &= (g_i - \sum_{k=0}^{N^2-1} |T_{i,k}| \cdot |G_k|) \cdot \frac{\eta_{max}}{2} \end{aligned} \quad (13)$$

Making use of Lemma 1 and selecting $g_i = a$ we get the lower bound of the $N \times N$ IDCT output from (13)

$$\hat{g}_i^{lower} \geq (a - \frac{|a+b|}{2} - N \cdot \frac{b-a}{2}) \cdot \frac{\eta_{max}}{2}$$

Based on the discussion above we conclude that

Theorem 1 For a $N \times N$ input block vector \vec{g} with its elements in the interval $[a, b]$ ($b > a$), after quantization and inverse quantization of DCT-coefficients with the quantization factor in the range $[0, \eta_{max}]$, the IDCT output is in the range

$$[(a - \frac{|a+b|}{2} - N \cdot \frac{b-a}{2}) \cdot \frac{\eta_{max}}{2}, (b + \frac{|a+b|}{2} + N \cdot \frac{b-a}{2}) \cdot \frac{\eta_{max}}{2}]$$

Obviously, the IDCT range of MPEG video coding is actually a special case of that we discussed above. In MPEG video coding [4][1] DCT and IDCT are carried out on the 8×8 blocks, a 8×8 block can be an intra-block or an inter-block. An intra-block has its pixel values ranged from 0 to 255, while the pixel values of an inter-block are in the interval $[-255, 255]$. What is more, a DCT-coefficient is first weighted with a weight from 1 to 255 and then quantized with a quantizer_scale_code ranged from 1 to 31, that lets it be possible for us to design a quantization matrix and to select a quantizer scale adequately so that quantization factor selections in (10) and (12) can be fulfilled.

Generally speaking, the quantizer used to quantize the DCT-coefficients of an intra-block (intra-quantizer) is different from that of an inter-block (inter-quantizer)[5][2], so is it with the inverse quantization of the DCT-coefficients [4][1]. The maximum quantization factor of an intra-block is only slightly (less than 10%) larger than that of an inter-block, but from theorem 1 we can obtain that under a same η_{max} the IDCT output range of an intra-block is $[-1147.5 \frac{\eta_{max}}{2}, 1402 \frac{\eta_{max}}{2}]$ while the IDCT output of an inter-block is in the range $[-2295 \frac{\eta_{max}}{2}, 2295 \frac{\eta_{max}}{2}]$, furthermore, in MPEG video coding the DC coefficient of an intra-block is represented with a precision from 8 to 11 bit. Hence, The IDCT output range of MPEG video coding is that of an inter-block, because the IDCT output range of an inter-block covers that of an intra-block.

The other problem is that there is a mismatch control and a saturation of the reconstructed DCT-coefficients before IDCT, the IDCT output range defined in Theorem 1 has not taken them into account.

In MPEG-1 video coding [4] the mismatch control is performed by a process $|\hat{G}_k| = |\hat{G}_k| - 1$ if a reconstructed DCT-coefficient ($\hat{G}_k \neq 0$) is even, this control makes the IDCT output range be a little bit narrower than that addressed in Theorem 1 as one can deduce from (11) and (13). In MPEG-2 video coding [1] the mismatch control only affects the last reconstructed DCT-coefficient \hat{G}_{63} and makes maximal ± 0.24 contribute to a reconstructed pixel \hat{g}_i , thus, this kind of influence can be ignored.

The saturation clips the reconstructed DCT-coefficients ($\vec{\hat{G}}$) to the range $[-2048, 2047]$. In MPEG-1 video coding the saturation is performed after the mismatch control, while it is carried out before the mismatch control in MPEG-2 coding. By substituting $N = 8$, $a = -255$ and $b = 255$ into Lemmea 2 we can directly obtain that the DCT-coefficients (\vec{G}) are in the interval $[-491, 491]$ when the IDCT output has its upper bound or lower bound, hence, the reconstructed DCT-coefficients should be never out of the range $[-2048, 2047]$ with the reasonable quantization and inverse quantization of DCT-coefficients. This proves that the saturation has no influence on the derived IDCT output range in Theorem 1. Therefore, the IDCT output range defined in Theorem 1 covers the range derived under consideration of the mismatch control and the saturation in MPEG video coding.

Based on the analysis above, we can directly deduce the IDCT output range of MPEG video coding from theorem 1 with $N = 8$ and $a = -255$ and $b = 255$ that

Corollary 1 For MPEG video coding where the pixel values of a 8×8 inter-block are in the interval $[-255, 255]$, after quantization and inverse quantization of DCT-coefficients with the quantization factor in the range $[0, \eta_{max}]$, the IDCT output is in the range

$$[-2295 \frac{\eta_{max}}{2}, 2295 \frac{\eta_{max}}{2}]$$

From corollary 1 we can acquire that for the MRE quantizer with $\eta_{max} = 2$ and for the Test Model quantizer with $\eta_{max} = 1.5$ used in [6] the IDCT output ranges are $[-2295, 2295]$ and $[-1721, 1721]$, respectively. They are the much smaller ranges compared with the ranges $[-12719, 12719]$ and $[-9539, 9539]$ derived in [6]. Note that in [6] Linzer derived the range $[1821, 12719]$ for the IDCT upper bound of the MRE quantizer and $[1586, 12719]$ for that of the Test Model quantizer, the upper bounds derived here lie in the ranges proposed in [6].

4 IDCT output range of MPEG video coding with TM5 quantizer

In MPEG video standards only the inverse quantization of a DCT-coefficient has been defined, this provides the free room for choosing the quantizers to quantize the DCT-coefficients at an encoder. One of these used quantizers is so called TM5 quantizer which is applied in the TM5 [2]. As mentioned in section 3, here we only need to discuss the TM5 inter-quantizer.

The maximum quantization factor of TM5 inter-quantizer is not a constant but is dependent on the amplitude of a DCT-coefficient. For a small DCT-coefficient in the range $[-7, 7]$ the maximum quantization factor can be as large as 3, while for a bigger DCT-coefficient with its absolute value larger than 105 the maximum quantization factor is less than 1.6. Therefore, we can not simply use Corollary 1 with $\eta_{max} = 3$ to compute the desired IDCT output range, as it can be too far from its real one. Hence, it is necessary to analyze the characteristic of TM5 inter-quantizer at first. Considering that the zero DCT-coefficients make no contribute to the reconstructed pixel, for the convenience of derivation we assume that the DCT-coefficients G_k appearing in the following formulas are non-zero coefficients.

Let G_k be an arbitrary DCT-coefficient of an 8×8 inter-block \vec{g} , Q be the quantizer scale, W_k be the k th element of the quantization matrix [4][1] with $1 \leq W_k \leq 255 \ \forall k = 0, 1, 2, \dots, 63$, the quantization of G_k in TM5 is defined as

$$K = (16 \cdot G_k) // W_k \quad (14)$$

$$K_q = K / Q \quad (15)$$

where K_q is the quantized DCT-coefficient of G_k . $//$ is the integer division with truncation of the result toward zero, while $'/'$ denotes the integer division with rounding to the nearest integer.

The inverse quantization of quantized DCT-coefficient K_q for an inter-block defined in [1] is same with that defined in [4], that is

$$\hat{G}_k = (2 \cdot K_q + \text{sign}(K_q)) \cdot Q \cdot W_k / 32 \quad (16)$$

where the function $\text{sign}(\cdot)$ is defined as

$$\text{sign}(K_q) = \begin{cases} 1 & \text{if } K_q > 0 \\ 0 & \text{if } K_q = 0 \\ -1 & \text{if } K_q < 0 \end{cases}$$

Now we will study the form of the quantization factor of a DCT-coefficient G_k of the TM5 inter-quantizer.

First, we study the case of a positive DCT-coefficient. If $G_k > 0$ formula (14) can be rewritten as

$$16 \cdot G_k = (K + \gamma) \cdot W_k \quad \text{with } -0.5 \leq \gamma < 0.5$$

Similarly, Formula (15) can be rewritten as

$$K = K_q \cdot Q + \delta \quad \text{with } 0 \leq \delta < Q$$

Therefore, we have

$$16 \cdot G_k = (K_q \cdot Q + \delta + \gamma) \cdot W_k \quad (17)$$

where $-0.5 \leq \gamma < 0.5$ and $0 \leq \delta < Q$.

The flexible selection of quantization matrix and quantizer scale ensures $K_q \neq 0$, from (16) and (17) we get the quantization factor of G_k

$$\eta_k = \frac{\hat{G}_k}{G_k} = \frac{16 \cdot (2 \cdot K_q + 1) \cdot Q / 32}{K_q \cdot Q + \delta + \gamma} \quad (18)$$

If $\delta = 0$, $\gamma = -0.5$ and $K_q = 1$, η_k in (18) has the maximum value. Therefore

$$\eta_k \leq 1.5 \frac{Q}{Q - 0.5} \quad (19)$$

Substituting $\delta = 0$, $\gamma = -0.5$ and $K_q = 1$ into (17) we have

$$Q = 0.5 + \frac{16 \cdot G_k}{W_k} \quad (20)$$

Replacing quantizer scale Q in (19) with the form of (20) we get

$$\eta_k \leq 1.5 \frac{32 \cdot G_k + W_k}{32 \cdot G_k} \quad (21)$$

Considering that the maximum value of W_k is 255, we have

$$\eta_k \leq 1.5 \cdot \left(1 + \frac{255}{32 \cdot G_k}\right) \text{ for } G_k > 0 \quad (22)$$

Similarly, for a negative DCT-coefficient G_k we have

$$\eta_k \leq 1.5 \cdot \left(1 - \frac{255}{32 \cdot G_k}\right) \text{ for } G_k < 0 \quad (23)$$

According to the analysis above, we conclude that

Theorem 2 *The quantization factor η_k of a non-zero DCT-coefficient G_k with the inter-quantization defined in TM5 [2] satisfies*

$$\eta_k \leq 1.5 \cdot \left(1 + \frac{255}{32 \cdot |G_k|}\right) \text{ for } G_k \neq 0, \quad k = 0, 1, 2, \dots, 63$$

With help of Theorem 2 we can be able to derive the IDCT output range of MPEG coding with TM5 quantizer. From (8) we have

$$\hat{g}_i = \sum_{k=0}^{63} T_{i,k} \cdot G_k \cdot \eta_k \leq \sum_{k=0}^{63} (T_{i,k} \cdot G_k + |T_{i,k} \cdot G_k|) \frac{\eta_k}{2} \quad (24)$$

Substituting η_k derived in Theorem 2 into (24) we get

$$\begin{aligned} \hat{g}_i &\leq \sum_{k=0}^{63} (T_{i,k} \cdot G_k + |T_{i,k} \cdot G_k|) \cdot 0.75 \cdot \left(1 + \frac{255}{32 \cdot |G_k|}\right) \\ &\leq 0.75 \cdot (g_i + \sum_{k=0}^{63} |T_{i,k}| |G_{i,k}|) + 0.75 \cdot \sum_{k=0}^{63} (T_{i,k} \cdot G_k + |T_{i,k} \cdot G_k|) \cdot \frac{255}{32 \cdot |G_k|} \end{aligned}$$

It should be noted here that the sum $0.75 \cdot (g_i + \sum_{k=0}^{63} |T_{i,k}| |G_{i,k}|)$ is actually the upper bound of IDCT output of MPEG video coding with a constant maximum quantization factor $\eta_{max} = 1.5$. Taking advantage of corollary 1 and considering $T_{i,k} \cdot G_k \leq |T_{i,k} \cdot G_k|$ we get the upper bound of \hat{g}_i

$$\hat{g}_i^{upper} \leq 0.75 \cdot (2295 + \frac{255}{16} \sum_{k=0}^{63} |T_{i,k}|) \quad (25)$$

For 8×8 IDCT one can exactly calculate the sum

$$\sum_{k=0}^{63} |T_{i,k}| = 6.97935 \quad (26)$$

From (25) and (26) we obtain

$$\hat{g}_i^{upper} \leq 1805$$

In the same way we can get the lower bound of \hat{g}_i

$$\hat{g}_i^{lower} \geq -1805$$

Therefore, we have

Theorem 3 *For MPEG video coding where the pixel values of a 8×8 inter-block are in the interval $[-255, 255]$, after the quantization of DCT-coefficients defined TM5 [2] and the standard MPEG inverse quantization of DCT-coefficients [1] [4], the IDCT output has a range of*

$$[-1805, 1805]$$

The inter-quantizer defined in MPEG-1 test model SM3 [5] is slightly different from the TM5 quantizer discussed above, but through the similar analysis we can obtain that the IDCT output range should be $[-1972, 1972]$ for MPEG video coding with the quantization of DCT-coefficients defined in SM3 [5].

As the derived range $[-1805, 1805]$ is within the range $[-2048, 2047]$, 12 bit should be sufficient to represent the IDCT output before clipping for MPEG video coding with TM5 quantizer, this is also valid for MPEG video coding with SM3 quantizer.

5 IDCT output with TM5 quantizer

This section provides some examples to verify the theoretical analysis in section 3 and 4. The quantizer used here is TM5 quantizer, the mismatch control and saturation defined in MPEG-2 video standard is also taken into account.

In the first example we expect that the input block vector can make the IDCT output approaching the IDCT upper bound. In particular, we obtain an input block vector

$$\begin{pmatrix} 255 & 255 & 255 & -255 & -255 & 255 & 255 & 255 \\ -255 & -255 & -255 & -255 & -255 & 255 & 255 & -255 \\ -255 & -255 & -255 & -255 & -255 & 255 & 255 & -255 \\ 255 & -255 & -255 & -255 & -255 & 255 & 255 & 255 \\ 255 & -255 & -255 & -255 & -255 & 255 & 255 & 255 \\ 255 & 255 & 255 & 255 & 255 & -255 & -255 & 255 \\ 255 & 255 & 255 & 255 & 255 & -255 & -255 & 255 \\ -255 & 255 & 255 & 255 & 255 & 255 & 255 & -255 \end{pmatrix}$$

a quantization matrix

$$\begin{pmatrix} 129 & 255 & 169 & 107 & 255 & 72 & 70 & 255 \\ 255 & 255 & 234 & 211 & 255 & 141 & 97 & 255 \\ 169 & 166 & 255 & 255 & 255 & 255 & 255 & 33 \\ 215 & 211 & 199 & 255 & 215 & 255 & 82 & 42 \\ 129 & 255 & 169 & 107 & 255 & 72 & 70 & 255 \\ 143 & 141 & 133 & 255 & 143 & 255 & 54 & 27 \\ 70 & 68 & 255 & 255 & 255 & 255 & 255 & 13 \\ 255 & 255 & 46 & 42 & 255 & 27 & 19 & 255 \end{pmatrix}$$

and a *quantizer_scale* = 32 (i.e. *quantizer_scale_code* = 16, *q_scale_type* = 0) which result in an IDCT output

$$\begin{pmatrix} 1711 & 195 & 193 & -113 & -115 & 193 & 194 & 387 \\ -192 & -196 & -192 & -194 & -193 & 194 & 194 & -194 \\ -196 & -194 & -195 & -193 & -193 & 195 & 193 & -194 \\ 193 & -194 & -194 & -194 & -197 & 193 & 195 & 194 \\ 193 & -194 & -193 & -194 & -192 & 193 & 193 & 195 \\ 195 & 193 & 194 & 194 & 193 & -194 & -193 & 195 \\ 193 & 196 & 193 & 193 & 193 & -195 & -194 & 193 \\ -388 & 193 & 195 & 114 & 113 & 194 & 193 & -160 \end{pmatrix}$$

The maximum value of the IDCT output is 1711 which is quite near the upper bound 1805. In the second example we use

$$\begin{pmatrix} -255 & -255 & -255 & 255 & 255 & -255 & -255 & -255 \\ 255 & 255 & 255 & 255 & 255 & -255 & -255 & 255 \\ 255 & 255 & 255 & 255 & 255 & -255 & -255 & 255 \\ -255 & 255 & 255 & 255 & 255 & -255 & -255 & -255 \\ -255 & 255 & 255 & 255 & 255 & -255 & -255 & -255 \\ -255 & -255 & -255 & -255 & -255 & 255 & 255 & -255 \\ -255 & -255 & -255 & -255 & -255 & 255 & 255 & -255 \\ 255 & -255 & -255 & -255 & -255 & -255 & -255 & 255 \end{pmatrix}$$

as the input block vector which is exactly the inverse input vector used in the first example, and the same quantization matrix and quantizer scale applied in the first example, we obtain an IDCT output

$$\begin{pmatrix} -1711 & -195 & -193 & 113 & 115 & -193 & -194 & -387 \\ 192 & 196 & 192 & 194 & 193 & -194 & -194 & 194 \\ 196 & 194 & 195 & 193 & 193 & -195 & -193 & 194 \\ -193 & 194 & 194 & 194 & 197 & -193 & -195 & -194 \\ -193 & 194 & 193 & 194 & 192 & -193 & -193 & -195 \\ -195 & -193 & -194 & -194 & -193 & 194 & 193 & -195 \\ -193 & -196 & -193 & -193 & -193 & 195 & 194 & -193 \\ 388 & -193 & -195 & -114 & -113 & -194 & -193 & 160 \end{pmatrix}$$

which has the minimum value of -1711 and is precisely the inverse IDCT output vector of the first example.

These two examples have proved that the IDCT output can be extremely so large as declared in Theorem 3. But in the real cases, we would never use a quantization matrix like those applied in the two examples, and the input block would not be such a special one, therefore, it would be useful to study the IDCT output by using the realistic quantization matrices and the reasonable input blocks.

It is difficult to analyze under what kind of condition the IDCT output has the maximum value with a given quantization matrix and how large the maximum IDCT value can be. Nevertheless, we can find those meaningful examples to find out the reasonable precision to represent IDCT output before clipping.

The first quantization matrix investigated is the default quantization matrix for an intra-block used in MPEG coding [1] [4], the DC value is represented by 8 bit (i.e *intra_DC_precision* = 0). The matrix

has a form of

$$\begin{pmatrix} 8 & 16 & 19 & 22 & 26 & 27 & 29 & 34 \\ 16 & 16 & 22 & 24 & 27 & 29 & 34 & 37 \\ 19 & 22 & 26 & 27 & 29 & 34 & 34 & 38 \\ 22 & 22 & 26 & 27 & 29 & 34 & 37 & 40 \\ 22 & 26 & 27 & 29 & 32 & 35 & 40 & 48 \\ 26 & 27 & 29 & 32 & 35 & 40 & 48 & 58 \\ 26 & 27 & 29 & 34 & 38 & 46 & 56 & 69 \\ 27 & 29 & 35 & 38 & 46 & 56 & 69 & 83 \end{pmatrix}$$

we use an input intra-block vector

$$\begin{pmatrix} 89 & 255 & 255 & 89 & 89 & 255 & 89 & 255 \\ 89 & 89 & 89 & 89 & 89 & 89 & 255 & 89 \\ 255 & 89 & 255 & 255 & 255 & 255 & 255 & 255 \\ 255 & 89 & 255 & 255 & 89 & 255 & 255 & 89 \\ 89 & 89 & 255 & 255 & 89 & 255 & 255 & 255 \\ 89 & 255 & 89 & 255 & 89 & 255 & 89 & 89 \\ 89 & 89 & 89 & 255 & 255 & 255 & 255 & 89 \\ 89 & 89 & 255 & 89 & 89 & 255 & 255 & 255 \end{pmatrix}$$

and a quantizer scale $quantizer_scale = 104$ (i.e. $quantizer_scale_code = 30$, $q_scale_type = 1$), which lead to an IDCT output with a maximum value of 536

$$\begin{pmatrix} 102 & 244 & 182 & 109 & 84 & 255 & 76 & 292 \\ 139 & 148 & 53 & 55 & 78 & 77 & 197 & 34 \\ 286 & 141 & 246 & 241 & 250 & 231 & \mathbf{536} & 208 \\ 251 & 62 & 346 & 227 & 67 & 161 & 286 & 139 \\ 73 & 54 & 205 & 222 & 64 & 277 & 248 & 258 \\ 100 & 232 & 48 & 210 & 129 & 234 & 190 & 95 \\ 56 & 164 & 81 & 181 & 228 & 261 & 302 & 173 \\ 74 & 37 & 295 & 135 & 182 & 189 & 223 & 184 \end{pmatrix}$$

The second quantization matrix we used is the one which is frequently applied in the MPEG video coding for the quantization of DCT-coefficients of an inter-block. It has the form

$$\begin{pmatrix} 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 \\ 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \\ 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 \\ 19 & 20 & 21 & 22 & 23 & 24 & 26 & 27 \\ 20 & 21 & 22 & 23 & 25 & 26 & 27 & 28 \\ 21 & 22 & 23 & 24 & 26 & 27 & 28 & 30 \\ 22 & 23 & 24 & 26 & 27 & 28 & 30 & 31 \\ 23 & 24 & 25 & 27 & 28 & 30 & 31 & 33 \end{pmatrix}$$

By using an input block

$$\begin{pmatrix} -128 & -128 & 128 & 128 & 128 & 128 & 128 & 128 \\ 128 & -128 & 128 & 128 & 128 & 128 & -128 & 128 \\ -128 & 128 & -128 & 128 & 128 & -128 & 128 & -128 \\ 128 & -128 & -128 & 128 & -128 & 128 & 128 & -128 \\ -128 & 128 & 128 & -128 & 128 & -128 & 128 & -128 \\ -128 & -128 & -128 & 128 & 128 & 128 & -128 & 128 \\ 128 & -128 & 128 & -128 & -128 & 128 & 128 & -128 \\ 128 & -128 & -128 & 128 & -128 & -128 & 128 & 128 \end{pmatrix}$$

and a quantizer scale $quantizer_scale = 112$ (i.e. $quantizer_scale_code = 31$ $q_scale_type = 1$) we have an IDCT output with a maximum value of 517

$$\begin{pmatrix} -86 & -131 & 163 & 128 & 139 & 186 & 29 & -14 \\ 70 & -60 & 132 & 123 & 49 & 192 & -190 & 62 \\ -108 & 68 & -96 & 98 & 83 & -163 & \mathbf{517} & -92 \\ 121 & -131 & -56 & 276 & -107 & 43 & 49 & 23 \\ -135 & 109 & 77 & -90 & 171 & -33 & 110 & -91 \\ -114 & -94 & -97 & 111 & 64 & 133 & -101 & 126 \\ 75 & -138 & 157 & -19 & -94 & 13 & 63 & -98 \\ 24 & -72 & -89 & -2 & -100 & -21 & 102 & 80 \end{pmatrix}$$

Seeing that the maximum IDCT value of those two realistic examples is in the interval $[512, 1023]$, we conclude that at least 11 bit should be used to represent the IDCT output before clipping and rounding of TM5 encoder in order to avoid IDCT overflow.

6 Conclusion

In this paper, the IDCT output range of MPEG video coding has been theoretically proved. Particularly, we have intensively analyzed the IDCT output range of MPEG video coding with TM5 quantizer, the theoretical derivation and simulation results show that with TM5 quantizer it needs at least 11 bit to represent the IDCT output before clipping to avoid IDCT overflow, and theoretically maximal 12 bit to cover all cases. The analysis and examples of this paper are a guide to the definition of the IDCT necessary precision in the MPEG video standard.

References

- [1] ISO/IEC IS 13818-2, ITU-T Recommendation H.262; *Generic coding of moving pictures and associated audio information*; Nov. 1994
- [2] "Test Model 5, Draft Revision 2", ISO/IEC JTC1/SC29/WG11/N0400, April 1993.
- [3] ISO/IEC DIS 13818-4, *Information technology - Generic coding of moving picture and associated audio information - Part 4: Compliance testing*, March, 1995
- [4] ISO/IEC 11172-2, *Information technology - Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s - Part 2: Video*, Jan. 1, 1993
- [5] "MPEG Video Simulation Model Three (SM3)", ISO/IEC JTC1/SC2/WG11/N0010, July 1990
- [6] Elliot Linzer, "Maximum IDCT output of MPEG encoders", ISO-IEC/JTC1/SC29/WG11 MPEG 95/265, July 19, 1995.
- [7] William K. Pratt, "Digital Image Processing", John & Sons, Inc. 1978.