|  |  |
| --- | --- |
| **Joint Collaborative Team on Video Coding (JCT-VC)**  **of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11**  5th Meeting: Geneva, 16-23 March, 2011 | Document: JCTVC-E333 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **Transform Dynamic Range Analysis** | | |
| *Status:* | Input Document to JCT-VC | | |
| *Purpose:* | Information | | |
| *Author(s) or Contact(s):* | Louis Kerofsky, Kiran Misra, Andrew Segall 5750 Pacific Rim Blvd Camas WA 98607 USA | Tel: Email: | +1 360 817 7644 [lkerofsky@sharplabs.com](mailto:lkerofsky@sharplabs.com)  [misrak@sharplabs.com](mailto:misrak@sharplabs.com)  [asegall@sharplabs.com](mailto:asegall@sharplabs.com) |
| *Source:* | Sharp Labs of America | | |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

This contribution determines bounds on the resource needs of the inverse transform operation used in HEVC. The analysis uses the HM-2.0 transform implementation for illustration but the methods are general. The analysis focuses on the bit-depth between transform stages necessary for inverse transform computation. The analysis uses a bound on the coefficient data prior to the transform to determine the intermediate bit-depth. Bounds assuming a linear transform are first derived. The transform used in HM-2.0 has nonlinearity due to conversion to integers at various stages. This error is analyzed and a bound on intermediate bit-depth increase modified suitably. For each transform size 4x4, 8x8, 16x16, and 32x32 bit-depth increase bounds are established. These bounds can be used to determine memory bit-depth as well as general dynamic range analysis.

# Introduction

A decoder implementation is faced with determining the necessary resources for worst case operation i.e. memory size, memory bit-depth, bandwidth, etc. We focus on the bit-depth demanded by the inverse transform definition at an intermediate stage between horizontal and vertical transforms. This bit-depth impacts the memory needed for the transpose buffer linking inverse transform stages. In HEVC [1], transform sizes of 4x4, 8x8, 16x16 and 32x32 are used. The larger transform sizes have the potential for significant bit-depth increase. The purpose of this analysis is to understand the potential resource demands of the transform. In H.264, a normative bit-exact transform definition is given. Unlike prior video coding standards, the transform calculation is precise in eliminating drift but removes some freedom of a decoder implementation. HEVC is following a similar goal of eliminating drift in transform calculations.

The bit-depth needed to support the worst case coefficient data may be significantly higher than the bit-depth needed to support coefficient data which results from a forward transform of data with a limited dynamic range. An illustration of this need for additional dynamic range support is shown in the figures below. The dynamic range of a simple rotation in 2-dimensions is examined with the dynamic range of input, coefficients, and reconstruction indicated. Figure 1 shows a region of limited dynamic range in two-dimensions. Figure 2 shows the result of a simple rotation on this input set. Note the DR necessary to support the coefficient data, shown by the yellow line, is greater than that needed to support the input even though the transform preserves the **L2** norm. Figure 3 shows the result of applying the inverse transform to the coefficient data. The dynamic range necessary to support the worst case coefficient data, shown by the red line, is significantly greater than the dynamic range necessary to support the reconstruction of transform results. In this simple example, the reconstruction of a worst case coefficient vector requires twice the dynamic range as reconstruction of forward transform coefficient data. This example indicates the importance of worst case dynamic range analysis.

|  |  |  |
| --- | --- | --- |
| Figure 1 Input | Figure 2 Coefficients | Figure 3 Reconstruction |

An important first step in determining the intermediate bit-depth support needed is bounding the size of the coefficient data input to the transform. The worst case input to a decoder is not the result of a forward transform since all forward transform data lies within a subset of the collection of coefficients that must be supported. For example, quantization may increase the dynamic range support needed at the decoder even when the coefficient data originates from a forward transform, for analysis showing the increased IDCT output range based following TM5 quantization with mpeg-2 see [2]. A decoder implementation is concerned with the DR support needed and may analyze the coefficient signaling allowed or use of limitations on the bitstream contents. As an example of the latter, H.264 [3] includes normative limits on the dynamic range at various points of the inverse transform calculation to guarantee 16-bit dynamic range. There are many statements of this form at various points in section 8 of H.264 for example the sentences below occur.

*A bitstream conforming to this Recommendation | International Standard shall not contain data that results in any element dcYij of dcY that exceeds the range of integer values from –215 to 215–1, inclusive.*

*A bitstream conforming to this Recommendation | International Standard shall not contain data that results in any element dcCij of dcC that exceeds the range of integer values from –215 to 215–1, inclusive.*

*The bitstream shall not contain data that results in any element dij of d with i, j = 0..3 that exceeds the range of integer values from –215 to 215–1, inclusive.*

*The bitstream shall not contain data that results in any element eij of e with i, j = 0..3 that exceeds the range of integer values from –215 to 215–1, inclusive.*

*The bitstream shall not contain data that results in any element fij of f with i, j = 0..3 that exceeds the range of integer values from –215 to 215–1, inclusive.*

*The bitstream shall not contain data that results in any element gij of g with i, j = 0..3 that exceeds the range of integer values from –215 to 215–1, inclusive.*

*The bitstream shall not contain data that results in any element hij of h with i, j = 0..3 that exceeds the range of integer values from –215 to 215–33, inclusive.* (The value 33 here is intentional to avoid 16-bit overflow in the addition of a rounding offset in the following line.)

To our knowledge the JM codec does not enforce these limits. Forward quantization and nearest neighbor quantization will meet the limits [4], but more sophisticated operations such as rate distortion optimal quantization have not been studied to our knowledge and in theory could violate these constraints without appropriate limitation.

In general the minimal DR needed to support reconstruction is difficult to determine particularly in high dimensions when the behavior of the forward transform and potential quantization must be accounted for. The analysis of this document assumes the magnitude of each coefficient is independently limited. This is summarized by stating a bound on the absolute value of the elements of each coefficient of the input vector to the inverse transform. This bound may depend upon block size in practice.

**Notation:** We describe notation for norms of a vector and matrix **M** below:



**Bounds for linear transforms**: Given a vector and a matrix **M**, bounds on the norm of the transform result, **M**, can be derived from bounds on the norm of the input vector and bounds on the norm of the matrix **M**. These bounds are summarized below. We remark that these bounds are tight in the sense that given a matrix **M**, there are input vectors, differing per bound, for which the bound is achieved. (A vector whose signs follow the signs of the row of M giving maximum **L1** norm or a vector with single element located at position where matrix **M** achieves its **L**∞ bound). From the norm definition,



two bound relations can be derived:



These bound relations can be established with appropriate application of the triangle inequality for norms and basic properties of maximum computations.



Results for a sample matrix multiply transform. A set of matrix multiply transforms has been proposed previously for use with HEVC[5]. These consist of a linear matrix multiply followed by a shift. The norm and shift values used for these matrices are listed in the table below. When the inverse transform is applied to the result of forward transform, the bit-depth range is limited. When the only limit on the input to the inverse transform is the **L**∞ bound on the coefficient vector, worst case input can increase the bit depth in bit-depth. Typical coefficient data may be reduced in bit-depth for example a single DC coefficient will be spread among the block elements and hence reduced in magnitude.

|  |  |  |  |
| --- | --- | --- | --- |
| Size |  | Shift | Bit increase |
| 4x4 | 247 | 6 | 2 |
| 8x8 | 479 | 6 | 3 |
| 16x16 | 940 | 6 | 4 |
| 32x32 | 1862 | 6 | 5 |

From this data, we see that for worst case input, the bit increase ranges from 2 for a 4x4 transform to 5 for a 32x32 transform. Assuming a limit on coefficient input of 16-bits this design allows intermediates to exceed 16-bits. Limiting the 4x4 input to 14-bits would restrict the intermediate values to 16-bits likely without coding loss. Such a straightforward limitation for the 32x32 input would need to reduce the input to 11-bits. In this case, pre-clipping is unlikely to simultaneously blocking worst case input while maintaining coding performance.

**Extension of bounds to lifting based transforms i.e. HEVC**

The derivation of bounds assuming a linear transformation is straightforward as illustrated in the previous section. A lifting based transform as used in HEVC [1] is nonlinear due to loss of precision during right bit-shifts. Due to this nonlinearity, the previous results cannot apply directly. The prior bounds can be used to analyze a lifting transform which is represented as stages. The stages can be combined to derive a bound on the entire transform calculation.

Assumptions for analysis, valid for HEVC transform:

1. Transform is expressed as a concatenation of several stages.
2. At each stage, the output is the a linear matrix multiply followed by conversion to integer values which introduces an error of at most 1 in a subset of samples.

The stage decomposition allows analysis of independent stages. The error due only to truncation places a bound on the **L1** norm of the error between the lifting stage result and the result of a linear matrix.



This relation is true although we do not know the error vector but only a bound on its **L1** norm.

Concatenating these relations for all stages gives the equation



Denoting the product of the last k stage matrices by **Pk** we can simply this relation as follows:



An error bound can be determined from this relation as:



Using the two bounds above, the following bound holds:



This simplifies to



Observe that the bound has a term proportional to the norm of the input and a constant independent of the input signal. For large input signals, the constant term is irrelevant and can be accounted for by slightly increasing the gain of norm dependent term.

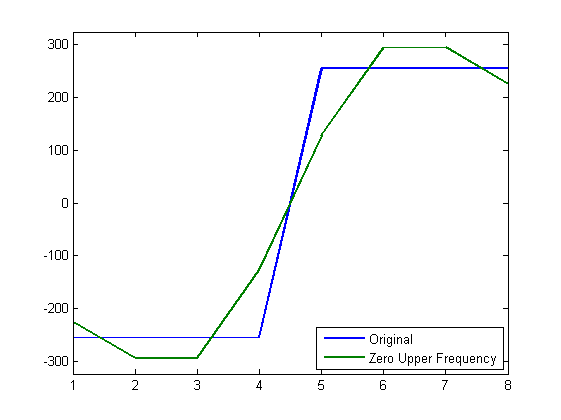
Determining bounds on the output is reduced to determining the norms of the product of stage matrices and counting the number of truncation errors in each stage to estimate the norm of the error vectors.

Results for various transform sizes based on code in HM-2.0.

|  |  |  |  |
| --- | --- | --- | --- |
| Size |  |  | Bit increase |
| 4x4 | 3.5 | 2 | 2 |
| 8x8 | 7.875 | 10 | 3 |
| 16x16 | 9.5638 | 34.1 | 4 |
| 32x32 | 18.8092 | 74.7 | 5 |

From this table, the HM2.0 transform will increase the intermediate values by no more than 2-bits for the 4x4 size and no more than 5-bits for the 32x32 size. For the HM2.0 codec, dynamic range expansion has been bounded by this analysis. We note that intermediate values exceeding 16-bits have been seen in tests of the codec. This is not unexpected as limiting intermediate bit-depth has not been a design goal.

**Example impact of quantization on reconstructed DR**. During this work we discovered an interesting aspect of the interaction between quantization and reconstructed dynamic range. Namely frequency quantization may decrease the energy of the reconstruction while also increasing the dynamic range of the reconstructed signal. This is in similar to the results analyzed more deeply in [2] for mpeg video coding and TM5 quantization. We illustrate this phenomenon with a simple example of an original signal =255\*[-1,-1,-1,-1, 1, 1, 1, 1] and reconstruction following zeroing upper 4 DCT components, a simple quantization method. The well-known ringing phenomenon is seen in this example. Clearly energy in the coefficient signal is reduced and therefore the energy of original signal is reduced since the DCT preserves energy. The figure below illustrates the increase in dynamic range needed to support the ringing of the reconstructed signal which requires a dynamic range approximately 14% larger than that of the original signal. The only thing particularly novel here is the observation that ringing increases the dynamic range of the reconstructed signal. [Note this simple coefficient domain process results in a coefficient vector which is not the forward transform of any signal with DR no more than 255].



The dynamic range increase illustrated by this single input signal and simple coefficient truncation process indicates the potential for dynamic range increase even though quantization reduces the energy. In general accounting for a range of input signals and various quantization operations which may occur due to rate-distortion optimization for example seem like challenging problems.

**Conclusions**: The methods presented here allow bounding the bit-depth expansion of linear or slightly non-linear transforms on worst case input. The analysis of worst case input to the inverse transform is an important aspect of this technique. This can be used during standard design to select appropriate limits used during transform calculation or by implementers to determine resources necessary to implement the transforms.

# Patent rights declaration(s)

**SHARP may have current or pending patent rights relating to the technology described in this contribution and, conditioned on reciprocity, is prepared to grant licenses under reasonable and non-discriminatory terms as necessary for implementation of the resulting ITU-T Recommendation | ISO/IEC International Standard (per box 2 of the ITU-T/ITU-R/ISO/IEC patent statement and licensing declaration form).**

**References**:

1. T. Wiegand, W.-J. Han, J.-R. Ohm, G. J. Sullivan, “High Efficiency Video Coding (HEVC) text specification Working Draft 1.” JCTVC-C403, Guangzhou, October, 2010.
2. M. Zhou, J. De Lameillieure, “IDCT output range before clipping of MPEG video coding” [Signal Processing: Image Communication](http://www.sciencedirect.com/science/journal/09235965) [Volume 11, Issue 2](http://www.sciencedirect.com/science?_ob=PublicationURL&_tockey=%23TOC%235640%231997%23999889997%2313728%23FLP%23&_cdi=5640&_pubType=J&view=c&_auth=y&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=26434d0ff173e07ed54a165e975ca0d1), December 1997, Pages 137-14
3. ITU-T Recommendation H.264 and ISO/IEC International Standard 14496-10 AVC
4. Louis Kerofsky, “Notes on JVT IDCT”, JVT-C24.doc, May 6-10, 2002
5. A.Fuldseth, G. Bjøntegaard, “Unified transform design for HEVC with 16 bit intermediate data representation”, JCTVC-D224 Daegu, KR, January 2011.