NETVC (Internet Video Codec)
Internet-Draft
Intended status: Informational
Expires: May 18, 2017

Y. Cho
Mozilla Corporation

November 14, 2016

# Applying PVQ Outside Daala
# draft-cho-netvc-applypvq-03

## Abstract

This document describes the Perceptual Vector Quantization (PVQ) outside of the Daala video codec, where PVQ was originally developed. It discusses the issues arising while integrating PVQ into a traditional video codec, AV1.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 18, 2017.

Table of Contents



## 1. Background

Perceptual Vector Quantization (PVQ) [Perceptual-VQ][I-D.valin-netvc-pvq] has been proposed as a quantization and coefficient coding tool for an internet video codec. PVQ was originally developed for the Daala video codec [1] [PVQ-demo], which does a gain-shape coding of transform coefficients instead of more traditional scalar quantization. (The original abbreviation of PVQ, "Pyramid Vector Quantizer", as in [I-D.valin-netvc-pvq] is now commonly expanded as "Perceptual Vector Quantization".)

The most distinguishing idea of PVQ is the way it references a predictor. With PVQ, we do not subtract the predictor from the input to produce a residual, which is then transformed and coded. Both the predictor and the input are transformed into the frequency domain. Then, PVQ applies a reflection to both the predictor and the input such that the prediction vector lies on one of the coordinate axes, and codes the angle between them. By not subtracting the predictor from the input, the gain of the predictor can be preserved and is explicitly coded, which is one of the benefits of PVQ. Since DC is not quantized by PVQ, the gain can be viewed as the amount of contrast in an image, which is an important perceptual parameter.

Also, an input block of transform coefficients is split into frequency bands based on their spatial orientation and scale. Then, each band is quantized by PVQ separately. The 'gain' of a band indicates the amount of contrast in the corresponding orientation and scale. It is simply the L2 norm of the band. The gain is non-linearly companded and then scalar quantized and coded. The

remaining information in the band, the 'shape', is then defined as a point on the surface of a unit hypersphere.

Another benefit of PVQ is activity masking based on the gain, which automatically controls the quantization resolution based on the image contrast without any signaling. For example, for a smooth image area (i.e. low contrast thus low gain), the resolution of quantization will increase, thus fewer quantization errors will be shown. A succinct summary on the benefits of PVQ can be found in the Section 2.4 of [Terriberry_16].

Since PVQ has only been used in the Daala video codec, which contains many non-traditional design elements, there has not been any chance to see the relative coding performance of PVQ compared to scalar quantization in a more traditional codec design. We have tried to apply PVQ in the AV1 video codec, which is currently being developed by Alliance for Open Media (AOM) as an open source and royalty-free video codec. While most of benefits of using PVQ arise from the enhancement of subjective quality of video, compression results with activity masking enabled are not available yet in this draft because the required parameters, which were set for Daala, have not been adjusted to AV1 yet. These results were achieved optimizing solely for PSNR.

## 2. Integration of PVQ into non-Daala codec, AV1

Adopting PVQ in AV1 requires replacing both the scalar quantization step and the coefficient coding of AV1 with those of PVQ. In terms of inputs to PVQ and the usage of transforms as shown in Figure 1 and Figure 2, the biggest conceptual changes required in a traditional coding system, such as AV1, are

- Introduction of a transformed predictor both in encoder and decoder. For this, we apply a forward transform to the predictors, both intra-predicted pixels and inter-predicted (i.e., motion-compensated) pixels. This is because PVQ references the predictor in the transform domain, instead of using a pixel-domain residual as in traditional scalar quantization.

- Absence of a difference signal (i.e. residue) defined as "input source - predictor". Hence AV1 with PVQ does not do any 'subtraction' in order for an input to reference the predictor. Instead, PVQ takes a different approach to referencing the predictor which happens in the transform domain.

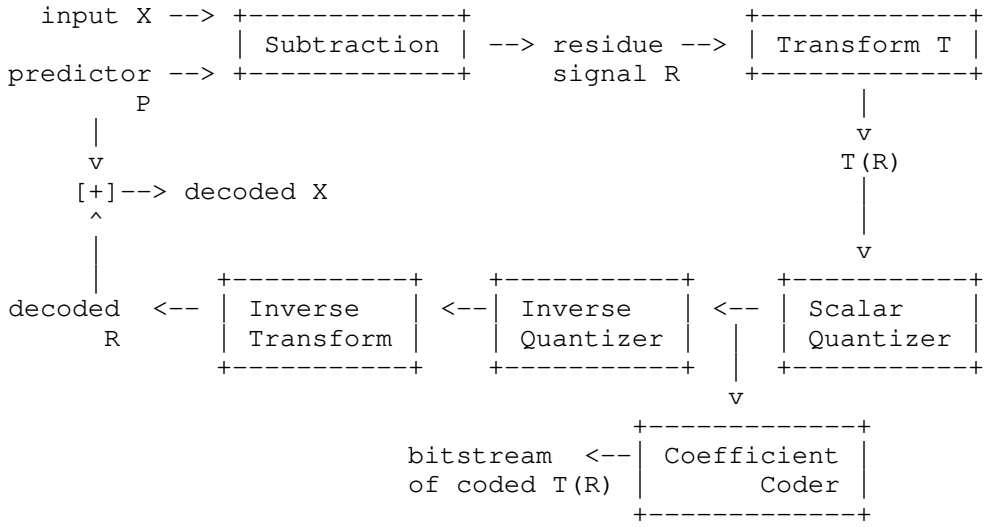



Figure 1: Traditional architecture containing Quantization and Transforms

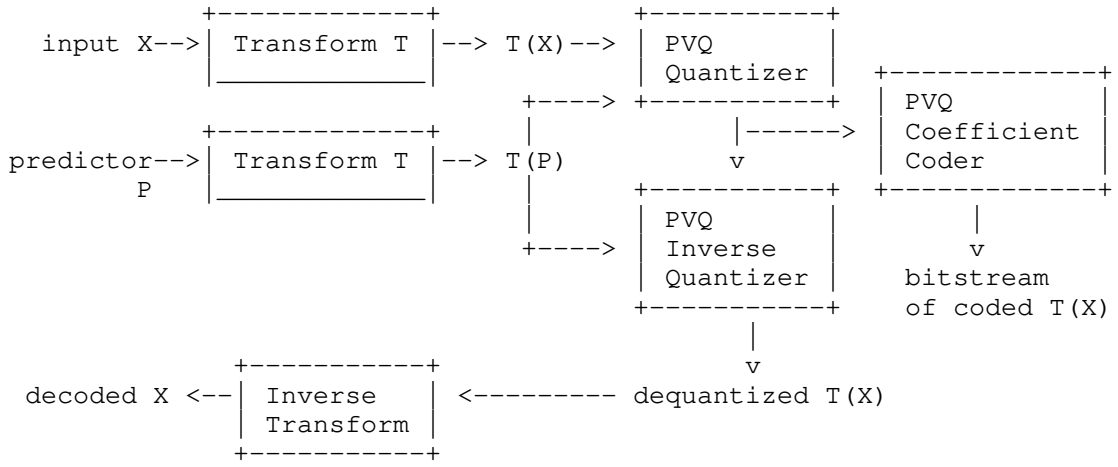



Figure 2: AV1 with PVQ

## 2.1. Signaling Skip for Partition and Transform Block

In AV1, a skip flag for a partition block is true if all of the quantized coefficients in the partition are zeros. The signaling for the prediction mode in a partition cannot be skipped. If the skip flag is true with PVQ, the predicted pixels are the final decoded

pixels (plus frame wise in-loop filtering such as deblocking) as in AV1 then a forward transform of a predictor is not required.

While AV1 currently defines only one 'skip' flag for each 'partition' (a unit where prediction is done), PVQ introduces another kind of 'skip' flag, called 'ac_dc_coded', which is defined for each transform block (and thus for each Y'CbCr plane as well). AV1 allows that a transform size can be smaller than a partition size which leads to partitions that can have multiple transform blocks. The ac_dc_coded flag signals whether DC and/or whole AC coefficients are coded by PVQ or not (PVQ does not quantize DC itself though).

### 2.2. Issues

- PVQ has its own rate-distortion optimization (RDO) that differs from that of traditional scalar quantization. This leads the balance of quality between luma and chroma to be different from that of scalar quantization. When scalar quantization of AV1 is done for a block of coefficients, RDO, such as trellis coding, can be optionally performed. The second pass of 2-pass encoding in AV1 currently uses trellis coding. When doing so it appears a different scaling factor is applied for each of Y'CbCr channels.

- In AV1, to optimize speed, there are inverse transforms that can skip applying certain 1D basis functions based on the distribution of quantized coefficients. However, this is mostly not possible with PVQ since the inverse transform is applied directly to a dequantized input, instead of a dequantized difference (i.e. input source - predictor) as in traditional video codec. This is true for both encoder and decoder.

- PVQ was originally designed for the 2D DCT, while AV1 also uses a hybrid 2D transform consisting of a 1D DCT and a 1D ADST. This requires PVQ to have new coefficient scanning orders for the two new 2D transforms, DCT-ADST and ADST-DCT (ADST-ADST uses the same scan order as for DCT-DCT). Those new scan orders have been produced based on that of AV1, for each PVQ-defined-band of new 2D transforms.

## 3. Performance of PVQ in AV1

### 3.1. Coding Gain

With the encoding options specified by both NETVC ([2]) and AOM testing for the high latency case, PVQ gives similar coding efficiency to that of AV1, which is measured in PSNR BD-rate. Again, PVQ's activity masking is not turned on for this testing. Also,

scalar quantization has matured over decades, while video coding with PVQ is much more recent.

We compare the coding efficiency for the IETF test sequence set "objective-1-fast" defined in [3], which consists of sixteen of 1080p, seven of 720p, and seven of 640x360 sequences of various types of content, including slow/high motion of people and objects, animation, computer games and screen casting. The encoding is done for the first 30 frames of each sequence. The encoding options used is : "-end-usage=q -cq-level=x --passes=2 --good --cpu-used=0 --auto-alt-ref=2 --lag-in-frames=25 --limit=30", which is official test condition of IETF and AOM for high latency encoding except limiting 30 frames only.

For comparison reasons, some of the lambda values used in RDO are adjusted to match the balance of luma and chroma quality of the PVQ-enabled AV1 to that of current AV1.

- Use half the value of lambda during intra prediction for the chroma channels.
- Scale PVQ's lambda by 0.9 for the chroma channels.
- Do not do RDO of DC for the chroma channels.

The results are shown in Table 1, which is the BD-Rate change for several image quality metrics. (The encoders used to generate this result are available from the author's git repository [4] and AOM's repository [5].)

| Metric | AV1 --> AV1 with PVQ |
|---|---|
| PSNR | -0.17% |
| PSNR-HVS | 0.27% |
| SSIM | 0.93% |
| MS-SSIM | 0.14% |
| CIEDE2000 | -0.28% |

Table 1: Coding Gain by PVQ in AV1

### 3.2. Speed

Total encoding time increases roughly 20 times or more when intensive RDO options, such as "--passes=2 --good --cpu-used=0 --auto-alt-ref=2 --lag-in-frames=25", are enabled. The significant increase in encoding time is due to the increase of computation by the PVQ. The PVQ tries to find asymptotically-optimal codepoints (in RD optimization sense) on a hypersphere with a greedy search, which has time complexity close to O(n*n) for n coefficients. Meanwhile, scalar quantization has time complexity of O(n).

Compared to Daala, the search space for a RDO decision in AV1 is far larger because AV1 considers ten intra prediction modes and four different transforms (for the transform block sizes 4x4, 8x8, and 16x16 only), and the transform block size can be smaller than the prediction block size. Since the largest transform and the prediction sizes are currently 32x32 and 64x64 in AV1, PVQ can be called approximately 5,160 times more in AV1 than in Daala. Also, AV1 applies transform and quantization for each candidate of RDO.

As an example, AV1 calls the PVQ function 632,520 times to encode the grandma_qcif (176x144) clip in intra frame mode while Daala calls 3843 times only (for QP = 30 and 39 for AV1 and daala respectively, which corresponds to actual quantizer used for quantization being 38). So, PVQ was called 165 times more in AV1 than Daala.

Table 2 shows the frequency of function calls to PVQ and scalar quantizers in AV1 at each speed level (where AV1 encoding mode is 'good') for the same sequence and the QP as used in the above example. The first column indicates speed level, the second column shows the number of calls to PVQ's search inside each band (function pvq_search_rdo_double() in [6]), the third column shows the number of calls to PVQ quantization of a transform block (function od_pvq_encode() in [7]), and the fourth column shows the number of calls to AV1's block quantizer. Smaller speed level gives slower encoding but better quality for the same rate by doing more RDO optimizations. The major difference from speed level 4 to 3 is enabling a use of the transform block smaller than the prediction (i.e. partition) block.

| Speed Level | # of calls to AV1 quantizer | # of calls to PVQ quantizer | # of calls to PVQ search inside a band |
|---|---|---|---|
| 5 | 28,028 | 26,786 | 365,913 |
| 4 | 57,445 | 56,980 | 472,222 |
| 3 | 505,039 | 564,724 | 3,680,366 |
| 2 | 505,039 | 564,724 | 3,680,366 |
| 1 | 535,100 | 580,566 | 3,990,327 |
| 0 | 589,931 | 632,520 | 4,109,113 |

Table 2: Comparison of Frequency of Calls to PVQ and Scalar Quantizers in AV1

## 4. Future Work

Possible future work includes:

- Enable activity masking, which also needs a HVS-tuned quantization matrix (bandwise QP scalars).
- Adjust the balance between luma and chroma qualities, tuning for subjective quality.
- Optimize the speed of the PVQ code, adding SIMD.
- RDO with more model-driven decision making, instead of full transform + quantization.

## 5. Development Repository

The ongoing work of integrating PVQ into AV1 video codec is located at the git repository [8].

## 6. Acknowledgements

Thanks to Tim Terriberry for his proofreading and valuable comments. Also thanks to Guillaume Martres for his contributions to integrating PVQ into AV1 during his internship at Mozilla and Thomas Daede for providing and maintaining the testing infrastructure by way of the www.arewecompressedyet.com (AWCY) website. [9].

## 7. IANA Considerations

This memo includes no request to IANA.

## 8. References

### 8.1. Informative References

[I-D.valin-netvc-pvq]
Valin, J., "Pyramid Vector Quantization for Video Coding", draft-valin-netvc-pvq-00 (work in progress), June 2015.

[Perceptual-VQ]
Valin, JM. and TB. Terriberry, "Perceptual Vector Quantization for Video Coding", Proceedings of SPIE Visual Information Processing and Communication , February 2015, <https://arxiv.org/pdf/1602.05209v1.pdf>.

[PVQ-demo]
Valin, JM., "Daala: Perceptual Vector Quantization (PVQ)", November 2014, <https://people.xiph.org/~jm/daala/pvq_demo/>.

[Terriberry_16]
Terriberry, TB., "Perceptually-Driven Video Coding with the Daala Video Codec", Proceedings SPIE Volume 9971, Applications of Digital Image Processing XXXIX , September 2016, <https://arxiv.org/pdf/1610.02488.pdf>.

### 8.2. URIs

[1] https://xiph.org/daala/

[2] https://tools.ietf.org/html/draft-ietf-netvc-testing-03

[3] https://tools.ietf.org/html/draft-ietf-netvc-testing-03

[4] https://github.com/ycho/aom/commit/2478029a9b6d02ee2ccc9dbafe7809b5ef345814

[5] https://aomedia.googlesource.com/aom/+/59848c5c797ddb6051e88b283353c7562d3a2c24

[6] https://github.com/ycho/aom/blob/14981eebb4a08f74182cea3c17f7361bc79cf04f/av1/encoder/pvq_encoder.c#L84

[7] https://github.com/ycho/aom/blob/14981eebb4a08f74182cea3c17f7361bc79cf04f/av1/encoder/pvq_encoder.c#L763

[8] https://github.com/ycho/aom/tree/av1_pvq

[9] https://arewecompressedyet.com/

Author's Address

Yushin Cho
Mozilla Corporation
331 E. Evelyn Avenue
Mountain View, CA 94041
USA

Phone: +1 650 903 0800
Email: ycho@mozilla.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 4, 2017

A. Fuldseth
G. Bjontegaard
S. Midtskogen
T. Davies
M. Zanaty
Cisco

October 31, 2016

# Thor Video Codec
# draft-fuldseth-netvc-thor-03

## Abstract

This document provides a high-level description of the Thor video codec. Thor is designed to achieve high compression efficiency with moderate complexity, using the well-known hybrid video coding approach of motion-compensated prediction and transform coding.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 4, 2017.

## Copyright Notice

Table of Contents

## 1. Introduction

This document provides a high-level description of the Thor video codec. Thor is designed to achieve high compression efficiency with moderate complexity, using the well-known hybrid video coding approach of motion-compensated prediction and transform coding.

The Thor video codec is a block-based hybrid video codec similar in structure to widespread standards. The high level encoder and decoder structures are illustrated in Figure 1 and Figure 2 respectively.

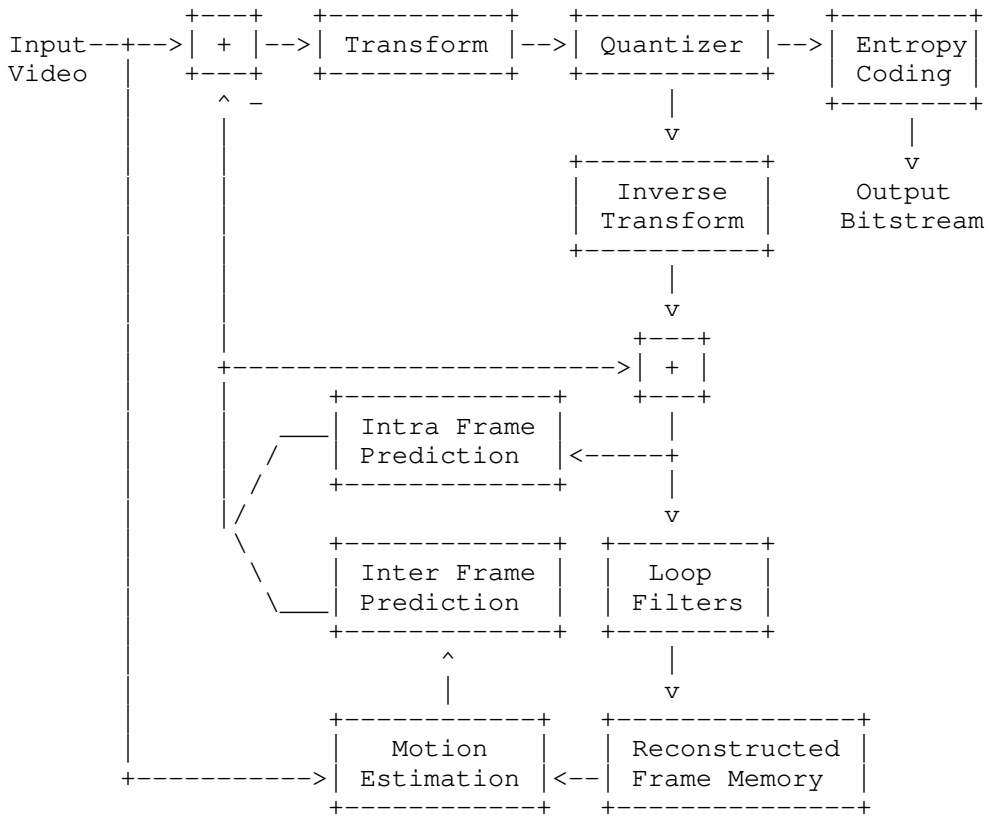



Figure 1: Encoder Structure

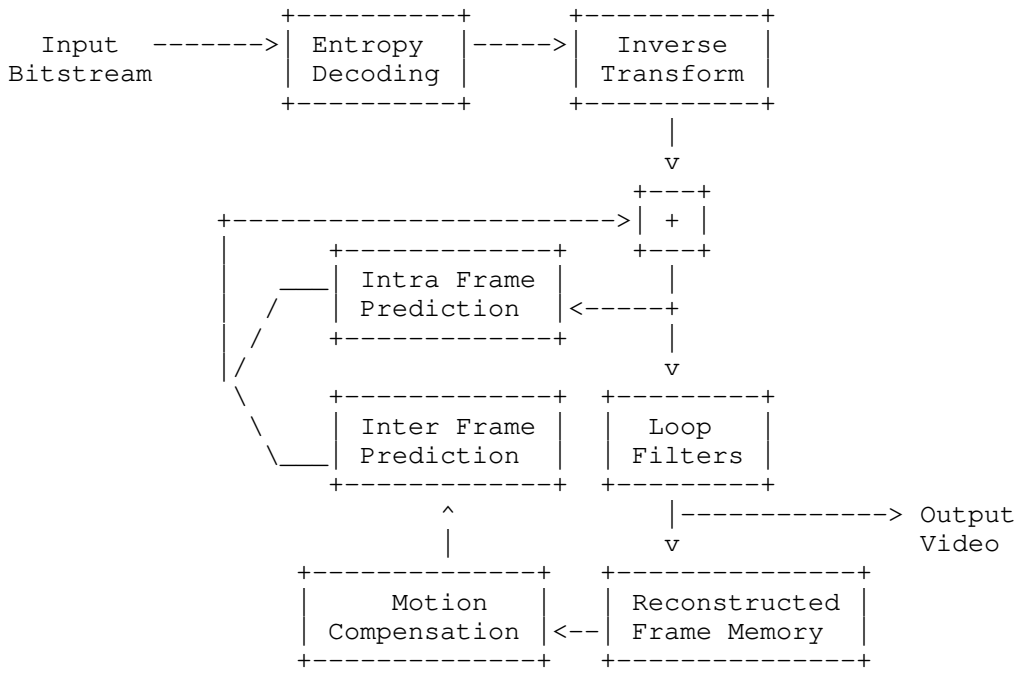



Figure 2: Decoder Structure

The remainder of this document is organized as follows. First, some requirements language and terms are defined. Block structures are described in detail, followed by intra-frame prediction techniques, inter-frame prediction techniques, transforms, quantization, loop filters, entropy coding, and finally high level syntax.

An open source reference implementation is available at github.com/cisco/thor.

## 2. Definitions

### 2.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

### 2.2. Terminology

This document frequently uses the following terms.

SB: Super Block - 64x64 or 128x128 block (luma pixels) which can be divided into CBs.

CB: Coding Block - Subdivision of a SB, down to 8x8 (luma pixels).

PB: Prediction Block - Subdivision of a CB, into 1, 2 or 4 equal blocks.

TB: Transform Block - Subdivision of a CB, into 1 or 4 equal blocks.

## 3. Block Structure

### 3.1. Super Blocks and Coding Blocks

Input frames with bitdepths of 8, 10 or 12 are supported. The internal bitdepth can be 8, 10 or 12 regardless if input bitdepth. The bitdepth of the output frames always follows the input frames. Chroma can be subsampled in both directions (4:2:0) or have full resolution (4:4:4).

Each frame is divided into 64x64 or 128x128 Super Blocks (SB) which are processed in raster-scan order. The SB size is signaled in the sequence header. Each SB can be divided into Coding Blocks (CB) using a quad-tree structure. The smallest allowed CB size is 8x8 luma pixels. The four CBs of a larger block are coded/signaled in the following order; upleft, downleft, upright, and downright.

The following modes are signaled at the CB level:

- Intra
- Inter0 (skip): MV index, no residual information
- Inter1 (merge): MV index, residual information
- Inter2 (uni-pred): explicit motion information, residual information
- Inter3 (ni-pred): explicit motion information, residual information

### 3.2. Special Processing at Frame Boundaries

At frame boundaries some square blocks might not be complete. For example, for 1920x1080 resolutions, the bottom row would consist of rectangular blocks of size 64x56. Rectangular blocks at frame boundaries are handled as follows. For each rectangular block, send one bit to choose between:

- A rectangular inter0 block and
- Further split.

For the bottom part of a 1920x1080 frame, this implies the following:

- For each 64x56 block, transmit one bit to signal a 64x56 inter0 block or a split into two 32x32 blocks and two 32x24 blocks.
- For each 32x24 block, transmit one bit to signal a 32x24 inter0 block or a split into two 16x16 blocks and two 16x8 blocks.
- For each 16x8 block, transmit one bit to signal a 16x8 inter0 block or a split into two 8x8 blocks.

Two examples of handling 64x56 blocks at the bottom row of a 1920x1080 frame are shown in Figure 3 and Figure 4 respectively.

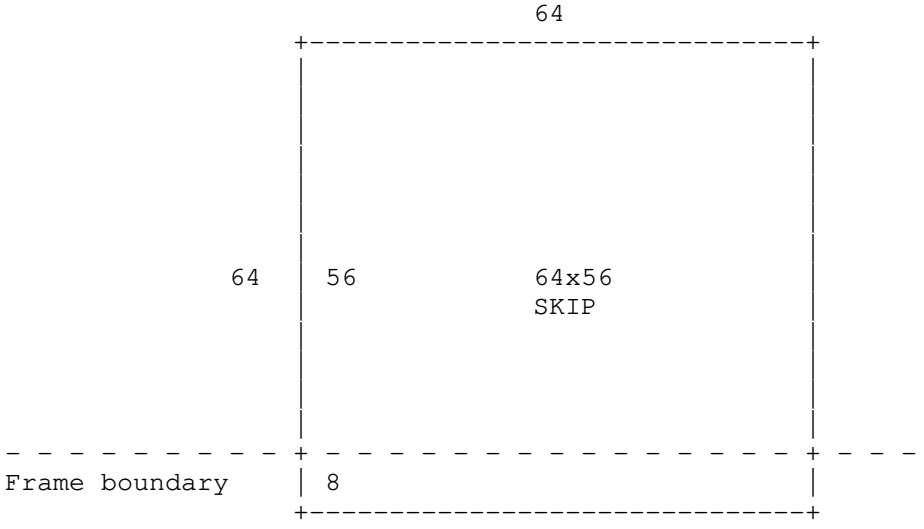



Figure 3: Super block at frame boundary

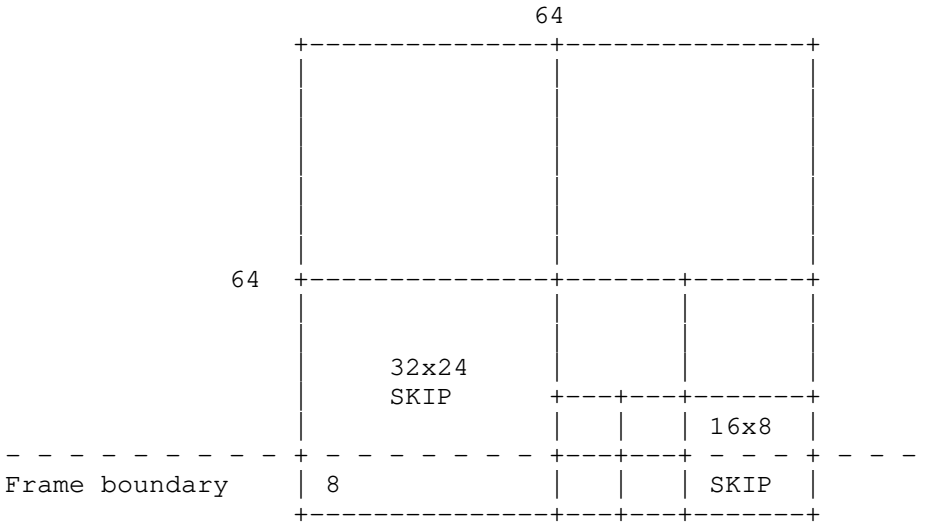



Figure 4: Coding block at frame boundary

## 3.3. Transform Blocks

A coding block (CB) can be divided into four smaller transform blocks (TBs).

## 3.4. Prediction Blocks

A coding block (CB) can also be divided into smaller prediction blocks (PBs) for the purpose of motion-compensated prediction. Horizontal, vertical and quad split are used.

# 4. Intra Prediction

8 intra prediction modes are used:

1. DC
2. Vertical (V)
3. Horizontal (H)
4. Upupright (north-northeast)
5. Upupleft (north-northwest)
6. Upleft (northwest)

7. Upleftleft (west-northwest)

8. Downleftleft (west-southwest)

The definition of DC, vertical, and horizontal modes are straightforward.

The upleft direction is exactly 45 degrees.

The upupright, upupleft, and upleftleft directions are equal to arctan(1/2) from the horizontal or vertical direction, since they are defined by going one pixel horizontally and two pixels vertically (or vice versa).

For the 5 angular intra modes (i.e. angle different from 90 degrees), the pixels of the neighbor blocks are filtered before they are used for prediction:

$$y(n) = (x(n-1) + 2*x(n) + x(n+1) + 2)/4$$

For the angular intra modes that are not 45 degrees, the prediction sometimes requires sample values at a half-pixel position. These sample values are determined by an additional filter:

$$z(n + 1/2) = (y(n) + y(n+1))/2$$

## 5. Inter Prediction

### 5.1. Multiple Reference Frames

Multiple reference frames are currently implemented as follows.

- Use a sliding-window process to keep the N most recent reconstructed frames in memory. The value of N is signaled in the sequence header.
- In the frame header, signal which of these frames shall be active for the current frame.
- For each CB, signal which of the active frames to be used for MC.

Combined with re-ordering, this allows for MPEG-1 style B frames.

A desirable future extension is to allow long-term reference frames in addition to the short-term reference frames defined by the sliding-window process.

### 5.2. Bi-Prediction

In case of bi-prediction, two reference indices and two motion vectors are signaled per CB. In the current version, PB-split is not allowed in bi-prediction mode. Sub-pixel interpolation is performed for each motion vector/reference index separately before doing an average between the two predicted blocks:

$$p(x,y) = (p0(x,y) + p1(x,y))/2$$

### 5.3. Improved chroma prediction

If specified in the sequence header, the chroma prediction, both intra and inter, or either, is improved by using the luma reconstruction if certain criteria are met. The process is described in the separate CLPF draft [I-D.midtskogen-netvc-chromapred].

### 5.4. Reordered Frames

Frames may be transmitted out of order. Reference frames are selected from the sliding window buffer as normal.

### 5.5. Interpolated Reference Frames

A flag is sent in the sequence header indicating that interpolated reference frames may be used.

If a frame is using an interpolated reference frame, it will be the first reference in the reference list, and will be interpolated from the second and third reference in the list. It is indicated by a reference index of -1 and has a frame number equal to that of the current frame.

The interpolated reference is created by a deterministic process common to the encoder and decoder, and described in the separate IRFVC draft [I-D.davies-netvc-irfvc].

### 5.6. Sub-Pixel Interpolation

#### 5.6.1. Luma Poly-phase Filter

Inter prediction uses traditional block-based motion compensated prediction with quarter pixel resolution. A separable 6-tap poly-phase filter is the basis method for doing MC with sub-pixel accuracy. The luma filter coefficients are as follows:

When bi-prediction is enabled in the sequence header:

1/4 phase: [2,-10,59,17,-5,1]/64

2/4 phase: [1,-8,39,39,-8,1]/64

3/4 phase: [1,-5,17,59,-10,2]/64

When bi-prediction is disabled in the sequence header:

1/4 phase: [1,-7,55,19,-5,1]/64

2/4 phase: [1,-7,38,38,-7,1]/64

3/4 phase: [1,-5,19,55,-7,1]/64

With reference to Figure 5, a fractional sample value, e.g. i0,0 which has a phase of 1/4 in the horizontal dimension and a phase of 1/2 in the vertical dimension is calculated as follows:

$a_{0,j} = 2*A_{-2,i} - 10*A_{-1,i} + 59*A_{0,i} + 17*A_{1,i} - 5*A_{2,i} + 1*A_{3,i}$

where $j = -2,...,3$

$i_{0,0} = (1*a_{0,-2} - 8*a_{0,-1} + 39*a_{0,0} + 39*a_{0,1} - 8*a_{0,2} + 1*a_{0,3} + 2048)/4096$

The minimum sub-block size is 8x8.

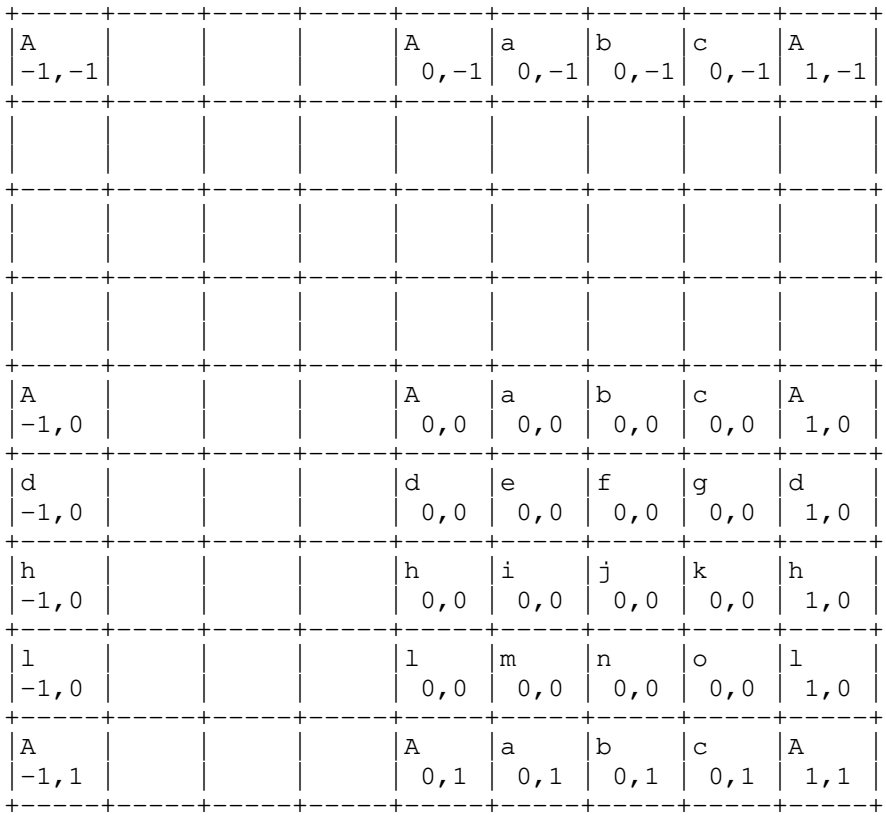



Figure 5: Sub-pixel positions

5.6.2. Luma Special Filter Position

For the fractional pixel position having exactly 2 quarter pixel offsets in each dimension, a non-separable filter is used to calculate the interpolated value. With reference to Figure 5, the center position j0,0 is calculated as follows:

j0,0 =

[0*A-1,-1 + 1*A0,-1 + 1*A1,-1 + 0*A2,-1 +

1*A-1,0 + 2*A0,0 + 2*A1,0 + 1*A2,0 +

1*A-1,1 + 2*A0,1 + 2*A1,1 + 1*A2,1 +

0*A-1,2 + 1*A0,2 + 1*A1,2 + 0*A2,2 + 8]/16

#### 5.6.3. Chroma Poly-phase Filter

Chroma interpolation is performed with 1/8 pixel resolution using the following poly-phase filter.

1/8 phase: [-2, 58, 10, -2]/64

2/8 phase: [-4, 54, 16, -2]/64

3/8 phase: [-4, 44, 28, -4]/64

4/8 phase: [-4, 36, 36, -4]/64

5/8 phase: [-4, 28, 44, -4]/64

6/8 phase: [-2, 16, 54, -4]/64

7/8 phase: [-2, 10, 58, -2]/64

### 5.7. Motion Vector Coding

#### 5.7.1. Inter0 and Inter1 Modes

Inter0 and inter1 modes imply signaling of a motion vector index to choose a motion vector from a list of candidate motion vectors with associated reference frame index. A list of motion vector candidates are derived from at most two different neighbor blocks, each having a unique motion vector/reference frame index. Signaling of the motion vector index uses 0 or 1 bit, dependent on the number of unique motion vector candidates. If the chosen neighbor block is coded in bi-prediction mode, the inter0 or inter1 block inherits both motion vectors, both reference indices and the bi-prediction property of the neighbor block.

For block sizes less than 64x64, inter0 has only one motion vector candidate, and its value is always zero.

Which neighbor blocks to use for motion vector candidates depends on the availability of the neighbor blocks (i.e. whether the neighbor blocks have already been coded, belong to the same slice and are not outside the frame boundaries). Four different availabilities, U, UR, L, and LL, are defined as illustrated in Figure 6. If the neighbor block is intra it is considered to be available but with a zero motion vector.

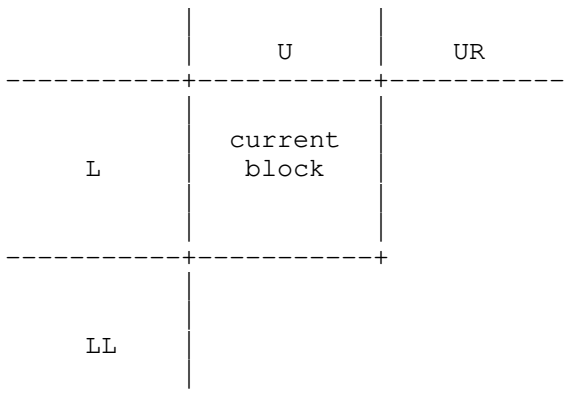



Figure 6: Availability of neighbor blocks

Based on the four availabilities defined above, each of the motion vector candidates is derived from one of the possible neighbor blocks defined in Figure 7.

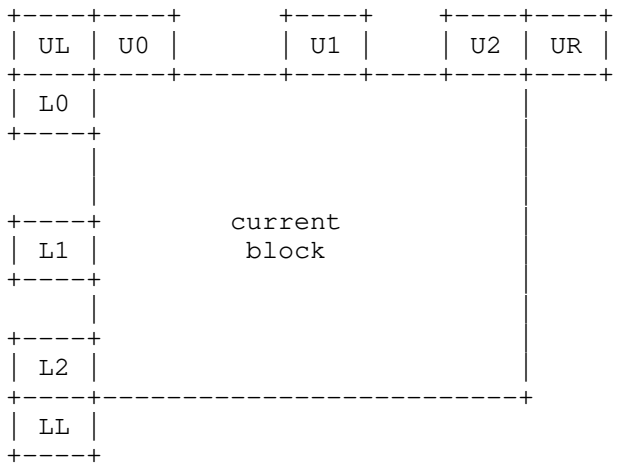



Figure 7: Motion vector candidates

The choice of motion vector candidates depends on the availability of neighbor blocks as shown in Table 1.

| U | UR | L | LL | Motion vector candidates |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | zero vector |
| 1 | 0 | 0 | 0 | U2, zero vector |
| 0 | 1 | 0 | 0 | NA |
| 1 | 1 | 0 | 0 | U2,zero vector |
| 0 | 0 | 1 | 0 | L2, zero vector |
| 1 | 0 | 1 | 0 | U2,L2 |
| 0 | 1 | 1 | 0 | NA |
| 1 | 1 | 1 | 0 | U2,L2 |
| 0 | 0 | 0 | 1 | NA |
| 1 | 0 | 0 | 1 | NA |
| 0 | 1 | 0 | 1 | NA |
| 1 | 1 | 0 | 1 | NA |
| 0 | 0 | 1 | 1 | L2, zero vector |
| 1 | 0 | 1 | 1 | U2,L2 |
| 0 | 1 | 1 | 1 | NA |
| 1 | 1 | 1 | 1 | U2,L2 |

Table 1: Motion vector candidates for different availability of neighbor blocks

### 5.7.2. Inter2 and Bi-Prediction Modes

Motion vectors are coded using motion vector prediction. The motion vector predictor is defined as the median of the motion vectors from three neighbor blocks. Definition of the motion vector predictor uses the same definition of availability and neighbors as in Figure 6 and Figure 7 respectively. The three vectors used for median filtering depends on the availability of neighbor blocks as shown in Table 2. If the neighbor block is coded in bi-prediction mode, only the first motion vector (in transmission order), MV0, is used as input to the median operator.

| U | UR | L | LL | Motion vectors for median filtering |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 3 x zero vector |
| 1 | 0 | 0 | 0 | U0,U1,U2 |
| 0 | 1 | 0 | 0 | NA |
| 1 | 1 | 0 | 0 | U0,U2,UR |
| 0 | 0 | 1 | 0 | L0,L1,L2 |
| 1 | 0 | 1 | 0 | UL,U2,L2 |
| 0 | 1 | 1 | 0 | NA |
| 1 | 1 | 1 | 0 | U0,UR,L2,L0 |
| 0 | 0 | 0 | 1 | NA |
| 1 | 0 | 0 | 1 | NA |
| 0 | 1 | 0 | 1 | NA |
| 1 | 1 | 0 | 1 | NA |
| 0 | 0 | 1 | 1 | L0,L2,LL |
| 1 | 0 | 1 | 1 | U2,L0,LL |
| 0 | 1 | 1 | 1 | NA |
| 1 | 1 | 1 | 1 | U0,UR,L0 |

Table 2: Neighbor blocks used to define motion vector predictor through median filtering

### 5.7.3. Motion Vector Direction

Motion vectors referring to reference frames later in time than the current frame are stored with their sign reversed, and these reversed values are used for coding and motion vector prediction.

## 6. Transforms

Transforms are applied at the TB or CB level, implying that transform sizes range from 4x4 to 128x128. The transforms form an embedded structure meaning the transform matrix elements of the smaller transforms can be extracted from the larger transforms.

## 7. Quantization

For the 32x32, 64x64 and 128x128 transform sizes, only the 16x16 low frequency coefficients are quantized and transmitted.

The 64x64 inverse transform is defined as a 32x32 transform followed by duplicating each output sample into a 2x2 block. The 128x128 inverse transform is defined as a 32x32 transform followed by duplicating each output sample into a 4x4 block.

### 7.1. Quantization matrices

A flag is transmitted in the sequence header to indicate whether quantization matrices are used. If this flag is true, a 6 bit value qmtx_offset is transmitted in the sequence header to indicate matrix strength.

If used, then in dequantization a separate scaling factor is applied to each coefficient, so that the dequantized value of a coefficient ci at position i is:

```
(ci * d(q) * IW(i,c,s,t,q) + 2^(k + 5)) >> (k + 6)
```

Figure 8: Equation 1

where IW is the scale factor for coefficient position i with size s, frame type (inter/inter) t, component (Y, Cb or Cr) c and quantizer q; and k=k(s,q) is the dequantization shift. IW has scale 64, that is, a weight value of 64 is no different to unweighted dequantization.

#### 7.1.1. Quantization matrix selection

The current luma qp value qpY and the offset value qmtx_offset determine a quantisation matrix set by the formula:

```
qmlevel = max(0,min(11,((qpY + qmtx_offset) * 12) / 44))
```

Figure 9: Equation 2

This selects one of the 12 different sets of default quantization matrix, with increasing qmlevel indicating increasing flatness.

For a given value of qmlevel, different weighting matrices are provided for all combinations of transform block size, type (intra/inter), and component (Y, Cb, Cr). Matrices at low qmlevel are flat (constant value 64). Matrices for inter frames have unity DC gain (i.e. value 64 at position 0), whereas those for intra frames are designed such that the inverse weighting matrix has unity energy gain (i.e. normalized sum-squared of the scaling factors is 1).

#### 7.1.2. Quantization matrix design

Further details on the quantization matrix and implementation can be found in the separate QMTX draft [I-D.davies-netvc-qmtx].

## 8. Loop Filtering

### 8.1. Deblocking

#### 8.1.1. Luma deblocking

Luma deblocking is performed on an 8x8 grid as follows:

1. For each vertical edge between two 8x8 blocks, calculate the following for each of line 2 and line 5 respectively:

   d = abs(a-b) + abs(c-d),

   where a and b, are on the left hand side of the block edge and c and d are on the right hand side of the block edge:

   a b | c d

2. For each line crossing the vertical edge, perform deblocking if and only if all of the following conditions are true:

   * d2+d5 < beta(QP)

   * The edge is also a transform block edge

   * abs(mvx(left)) > 2, or abs(mvx(right)) > 2, or

     abs(mvy(left)) > 2, or abs(mvy(right)) > 2, or

     One of the transform blocks on each side of the edge has non-zero coefficients, or

     One of the transform blocks on each side of the edge is coded using intra mode.

3. If deblocking is performed, calculate a delta value as follows:

   delta = clip((18*(c-b) - 6*(d-a) + 16)/32,tc,-tc),

   where tc is a QP-dependent value.

4. Next, modify two pixels on each side of the block edge as follows:

a' = a + delta/2

b' = b + delta

c' = c + delta

d' = d + delta/2

5. The same procedure is followed for horizontal block edges.

The relative positions of the samples, a, b, c, d and the motion vectors, MV, are illustrated in Figure 10.

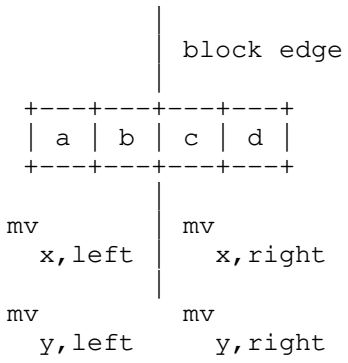



Figure 10: Deblocking filter pixel positions

8.1.2. Chroma Deblocking

Chroma deblocking is performed on a 4x4 grid as follows:

1. Delocking of the edge between two 4x4 blocks is performed if and only if:

   * The pixels on either side of the block edge belongs to an intra block.

   * The block edge is also an edge between two transform blocks.

2. If deblocking is performed, calculate a delta value as follows:

   delta = clip((4*(c-b) + (d-a) + 4)/8,tc,-tc),

   where tc is a QP-dependent value.

3. Next, modify one pixel on each side of the block edge as follows:

   b' = b + delta

   c' = c + delta

## 8.2. Constrained Low Pass Filter (CLPF)

A low-pass filter is applied after the deblocking filter if signaled in the sequence header. It can still be switched off for individual frames in the frame header. Also signaled in the frame header is whether to apply the filter for all qualified 128x128 blocks or to transmit a flag for each such block. A super block does not qualify if it only contains Inter0 (skip) coding block and no signal is transmitted for these blocks.

The filter is described in the separate CLPF draft [I-D.midtskogen-netvc-clpf].

# 9. Entropy coding

## 9.1. Overview

The following information is signaled at the sequence level:

- Sequence header

The following information is signaled at the frame level:

- Frame header

The following information is signaled at the CB level:

- Super-mode (mode, split, reference index for uni-prediction)
- Intra prediction mode
- PB-split (none, hor, ver, quad)
- TB-split (none or quad)
- Reference frame indices for bi-prediction
- Motion vector candidate index
- Transform coefficients if TB-split=0

The following information is signaled at the TB level:

- CBP (8 combinations of CBPY, CBPU, and CBPV)
- Transform coefficients

The following information is signaled at the PB level:

- Motion vector differences

## 9.2. Low Level Syntax

### 9.2.1. CB Level

```
    super-mode          (inter0/split/inter1/inter2-ref0/intra/inter2-ref1/inter
2-ref2/inter2-ref3,..)

    if (mode == inter0 || mode == inter1)

      mv_idx                    (one of up to 2 motion vector candidates)

    else if (mode == INTRA)

      intra_mode                (one of up to 8 intra modes)

      tb_split                  (NONE or QUAD, coded jointly with CBP for tb_spl
it=NONE)

    else if (mode == INTER)

      pb_split          (NONE,VER,HOR,QUAD)

      tb_split_and_cbp  (NONE or QUAD and CBP)

    else if (mode == BIPRED)

      mvd_x0, mvd_y0    (motion vector difference for first vector)

      mvd_x1, mvd_y1    (motion vector difference for second vector)

      ref_idx0, ref_idx1        (two reference indices)
```

### 9.2.2. PB Level

```
       if (mode == INTER2 || mode == BIPRED)

         mvd_x, mvd_y              (motion vector differences)
```

#### 9.2.3. TB Level

```
if (mode != INTER0 and tb_split == 1)

  cbp                        (8 possibilities for CBPY/CBPU/CBPV)

if (mode != INTER0)

  transform coefficients
```

#### 9.2.4. Super Mode

For each block of size NxN (64>=N>8), the following mutually exclusive events are jointly encoded using a single VLC code as follows (example using 4 reference frames):

If there is no interpolated reference frame:

```
INTER0      1
SPLIT       01
INTER1      001
INTER2-REF0 0001
BIPRED      00001
INTRA       000001
INTER2-REF1     0000001
INTER2-REF2     00000001
INTER2-REF3     00000000
```

If there is an interpolated reference frame:

```
INTER0      1
SPLIT       01
INTER1      001
BIPRED      0001
INTRA       00001
INTER2-REF1     000001
INTER2-REF2     0000001
INTER2-REF3     00000001
INTER2-REF0 00000000
```

If less than 4 reference frames is used, a shorter VLC table is used. If bi-pred is not possible, or split is not possible, they are omitted from the table and shorter codes are used for subsequent elements.

Additionally, depending on information from the blocks to the left and above (meta data and CBP), a different sorting of the events can be used, e.g.:

```
SPLIT       1
INTER1      01
INTER2-REF0 001
INTER0      0001
INTRA       00001
INTER2-REF1 000001
INTER2-REF2 0000001
INTER2-REF3 00000001
BIPRED      00000000
```

#### 9.2.5. CBP

Calculate code as follows:

```
if (tb-split == 0)

  N = 4*CBPV + 2*CBPU + CBPY

else

  N = 8
```

Map the value of N to code through a table lookup:

```
code = table[N]
```

where the purpose of the table lookup is the sort the different values of code according to decreasing probability (typically CBPY=1, CBPU=0, CBPV=0 having the highest probability).

Use a different table depending on the values of CBPY in neighbor blocks (left and above).

Encode the value of code using a systematic VLC code.

#### 9.2.6. Transform Coefficients

Transform coefficient coding uses a traditional zig-zag scan pattern to convert a 2D array of quantized transform coefficients, coeff, to a 1D array of samples. VLC coding of quantized transform coefficients starts from the low frequency end of the 1D array using two different modes; level-mode and run-mode, starting in level-mode:

- Level-mode
  - Encode each coefficient, coeff, separately
  - Each coefficient is encoded by:

+ The absolute value, level=abs(coeff), using a VLC code and

+ If level > 0, the sign bit (sign=0 or sign=1 for coeff>0 and coeff<0 respectively).

* If coefficient N is zero, switch to run-mode, starting from coefficient N+1.

o Run-mode

* For each non-zero coefficient, encode the combined event of:

1. Length of the zero-run, i.e. the number of zeros since the last non-zero coefficient.

2. Whether or not level=abs(coeff) is greater than 1.

3. End of block (EOB) indicating that there are no more non-zero coefficients.

* Additionally, if level = 1, code the sign bit.

* Additionally, if level > 1 define code = 2*(level-2)+sign,

* If the absolute value of coefficient N is larger than 1, switch to level-mode, starting from coefficient N+1.

Example

Figure 11 illustrates an example where 16 quantized transform coefficients are encoded.

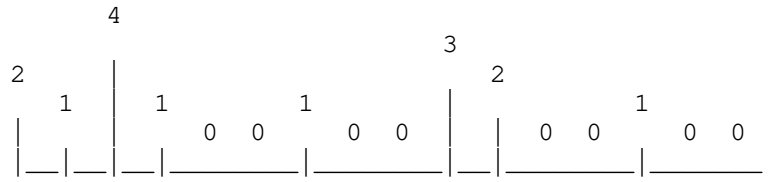



Figure 11: Coefficients to encode

Table 3 shows the mode, VLC number and symbols to be coded for each coefficient.

| Index | abs(coeff) | Mode | Encoded symbols |
|---|---|---|---|
| 0 | 2 | level-mode | level=2,sign |
| 1 | 1 | level-mode | level=1,sign |
| 2 | 4 | level-mode | level=4,sign |
| 3 | 1 | level-mode | level=1,sign |
| 4 | 0 | level-mode | level=0 |
| 5 | 0 | run-mode | |
| 6 | 1 | run-mode | (run=1,level=1) |
| 7 | 0 | run-mode | |
| 8 | 0 | run-mode | |
| 9 | 3 | run-mode | (run=1,level>1), 2*(3-2)+sign |
| 10 | 2 | level-mode | level=2, sign |
| 11 | 0 | level-mode | level=0 |
| 12 | 0 | run-mode | |
| 13 | 1 | run-mode | (run=1,level=1) |
| 14 | 0 | run-mode | EOB |
| 15 | 0 | run-mode | |

Table 3: Transform coefficient encoding for the example above.

## 10. High Level Syntax

High level syntax is currently very simple and rudimentary as the primary focus so far has been on compression performance. It is expected to evolve as functionality is added.

### 10.1. Sequence Header

- Width - 16 bits
- Height - 16 bits
- Enable/disable PB-split - 1 bit
- SB size - 3 bits
- Enable/disable TB-split - 1 bit
- Number of active reference frames (may go into frame header) - 2 bits (max 4)
- Enable/disable interpolated reference frames - 1 bit
- Enable/disable delta qp - 1 bit

- Enable/disable deblocking - 1 bit
- Constrained low-pass filter (CLPF) enable/disable - 1 bit
- Enable/disable block context coding - 1 bit
- Enable/disable bi-prediction - 1 bit
- Enable/disable quantization matrices - 1 bit
- If quantization matrices enabled: quantization matrix offset - 6 bits
- Select 420 or 444 input - 1 bit
- Number of reordered frames - 4 bits
- Enable/disable chroma intra prediction from luma - 1 bit
- Enable/disable chroma inter prediction from luma - 1 bit
- Internal frame bitdepth (8, 10 or 12 bits) - 2 bits
- Input video bitdepth (8, 10 or 12 bits) - 2 bits

### 10.2. Frame Header

- Frame type - 1 bit
- QP - 8 bits
- Identification of active reference frames - num_ref*4 bits
- Number of intra modes - 4 bits
- Number of active reference frames - 2 bits
- Active reference frames - number of active reference frames * 6 bits
- Frame number - 16 bits
- If CLPF is enabled in the sequence header: Constrained low-pass filter (CLPF) strength - 2 bits (00 = off, 01 = strength 1, 10 = strength 2, 11 = strength 4)
- IF CLPF is enabled in the sequence header: Enable/disable CLPF signal for each qualified filter block

## 11. IANA Considerations

This document has no IANA considerations yet. TBD

## 12. Security Considerations

This document has no security considerations yet. TBD

## 13. Normative References

[I-D.davies-netvc-irfvc]
Davies, T., "Interpolated reference frames for video coding", draft-davies-netvc-irfvc-00 (work in progress), October 2015.

[I-D.davies-netvc-qmtx]
Davies, T., "Quantisation matrices for Thor video coding", draft-davies-netvc-qmtx-00 (work in progress), March 2016.

[I-D.midtskogen-netvc-chromapred]
Midtskogen, S., "Improved chroma prediction", draft-midtskogen-netvc-chromapred-02 (work in progress), October 2016.

[I-D.midtskogen-netvc-clpf]
Midtskogen, S., Fuldseth, A., and M. Zanaty, "Constrained Low Pass Filter", draft-midtskogen-netvc-clpf-02 (work in progress), April 2016.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

## Authors' Addresses

Arild Fuldseth
Cisco
Lysaker
Norway

Email: arilfuld@cisco.com

Gisle Bjontegaard
Cisco
Lysaker
Norway

Email: gbjonteg@cisco.com

Steinar Midtskogen
Cisco
Lysaker
Norway

Email: stemidts@cisco.com

Thomas Davies
Cisco
London
UK

Email: thdavies@cisco.com

Mo Zanaty
Cisco
RTP,NC
USA

Email: mzanaty@cisco.com

Network Working Group
Internet Draft
Intended status: Informational

A. Filippov
Huawei Technologies
A. Norkin
Netflix
J.R. Alvarez
Huawei Technologies

Expires: April 20, 2020

November 21, 2019

# Video Codec Requirements and Evaluation Methodology
## draft-ietf-netvc-requirements-10.txt

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts can be accessed at http://datatracker.ietf.org/drafts/current/

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

This Internet-Draft will expire on April 21, 2020.

Abstract

This document provides requirements for a video codec designed mainly for use over the Internet. In addition, this document describes an evaluation methodology needed for measuring the compression efficiency to ensure whether the stated requirements are fulfilled or not.

Table of Contents

## 1. Introduction

In this document, the requirements for a video codec designed mainly for use over the Internet are presented. The requirements encompass a wide range of applications that use data transmission over the Internet including Internet video streaming, IPTV, peer-to-peer video conferencing, video sharing, screencasting, game streaming and video monitoring / surveillance. For each application, typical resolutions, frame-rates and picture access modes are presented. Specific requirements related to data transmission over packet-loss networks are considered as well. In this document, when we discuss data protection techniques we only refer to methods designed and implemented to protect data inside the video codec since there are many existing techniques that protect generic data transmitted over networks with packet losses. From the theoretical point of view, both packet-loss and bit-error robustness can be beneficial for video codecs. In practice, packet losses are a more significant problem than bit corruption in IP networks. It is worth noting that there is an evident interdependence between possible amount of delay and the necessity of error robust video streams:

- If an amount of delay is not crucial for an application, then reliable transport protocols such as TCP that retransmits undelivered packets can be used to guarantee correct decoding of transmitted data.

- If the amount of delay must be kept low, then either data transmission should be error free (e.g., by using managed networks) or compressed video stream should be error resilient.

Thus, error resilience can be useful for delay-critical applications to provide low delay in packet-loss environment.

## 2. Definitions and abbreviations used in this document

| Term | Meaning |
|---|---|
| High dynamic range imaging | is a set of techniques that allow a greater dynamic range of exposures or values (i.e., a wide range of values between light and dark areas) than normal digital imaging techniques. The intention is to accurately represent the wide range of intensity levels found in such examples as exterior scenes that include light-colored items struck by direct sunlight and areas of deep shadow [7]. |
| Random access period | is the period of time between two closest independently decodable frames (pictures). |
| RD-point | A point in a two dimensional rate-distortion space where the values of bitrate and quality metric are used as x- and y-coordinates, respectively |
| Visually lossless compression | is a form or manner of lossy compression where the data that are lost after the file is compressed and decompressed is not detectable to the eye; the compressed data appearing identical to the uncompressed data [8]. |
| Wide color gamut | is a certain complete color subset (e.g., considered in ITU-R BT.2020) that supports a wider range of colors (i.e., an extended range of colors that can be generated by a specific input or output device such as a video camera, monitor or printer and can be interpreted by a color model) than conventional color gamuts (e.g., considered in ITU-R BT.601 or BT.709). |

Table 1. Definitions used in the text of this document

| Abbreviation | Meaning |
|---|---|
| AI | All-Intra (each picture is intra-coded) |
| BD-Rate | Bjontegaard Delta Rate |
| FIZD | just the First picture is Intra-coded, Zero structural Delay |
| GOP | Group of Picture |
| HBR | High Bitrate Range |
| HDR | High Dynamic Range |
| HRD | Hypothetical Reference Decoder |
| IPTV | Internet Protocol Television |
| LBR | Low Bitrate Range |
| MBR | Medium Bitrate Range |
| MOS | Mean Opinion Score |
| MS-SSIM | Multi-Scale Structural Similarity quality index |
| PAM | Picture Access Mode |
| PSNR | Peak Signal-to-Noise Ratio |
| QoS | Quality of Service |
| QP | Quantization Parameter |
| RA | Random Access |
| RAP | Random Access Period |
| RD | Rate-Distortion |
| SEI | Supplemental Enhancement Information |
| UGC | User-Generated Content |
| VDI | Virtual Desktop Infrastructure |
| VUI | Video Usability Information |
| WCG | Wide Color Gamut |

Table 2. Abbreviations used in the text of this document

# 3. Applications

In this chapter, an overview of video codec applications that are currently available on the Internet market is presented. It is worth noting that there are different use cases for each application that define a target platform, and hence there are different types of communication channels involved (e.g., wired or wireless channels) that are characterized by different quality of service as well as bandwidth; for instance, wired channels are considerably more error-free than wireless channels and therefore require different QoS approaches. The target platform, the channel bandwidth and the channel quality determine resolutions, frame-rates and quality or bit-rates for video streams to be encoded or decoded. By default, color format YCbCr 4:2:0 is assumed for the application scenarios listed below.

## 3.1. Internet Video Streaming

Typical content for this application is movies, TV-series and shows, and animation. Internet video streaming uses a variety of client devices and has to operate under changing network conditions. For this reason, an adaptive streaming model has been widely adopted. Video material is encoded at different quality levels and different resolutions, which are then chosen by a client depending on its capabilities and current network bandwidth. An example combination of resolutions and bitrates is shown in Table 3.

A video encoding pipeline in on-demand Internet video streaming typically operates as follows:

- Video is encoded in the cloud by software encoders.
- Source video is split into chunks, each of which is encoded separately, in parallel.
- Closed-GOP encoding with 2-5 second intra-picture intervals (or more) is used.
- Encoding is perceptually optimized. Perceptual quality is important and should be considered during the codec development.

| Resolution * | Frame-rate, fps | PAM |
|---|---|---|
| 4K, 3840x2160 | 24/1.001, 24, 25, | RA |
| 2K (1080p), 1920x1080 | 30/1.001, 30, 50, | RA |
| 1080i, 1920x1080* | 60/1.001, 60, 100, | RA |
| 720p, 1280x720 | 120/1.001, 120 | RA |
| 576p (EDTV), 720x576 | The set of frame-rates | RA |
| 576i (SDTV), 720x576* | presented in this table | RA |
| 480p (EDTV), 720x480 | is taken from Table 2 | RA |
| 480i (SDTV), 720x480* | in [1] | RA |
| 512x384 | | RA |
| QVGA, 320x240 | | RA |

Table 3. Internet Video Streaming: typical values of resolutions, frame-rates, and RAPs

NB *: Interlaced content can be handled at the higher system level and not necessarily by using specialized video coding tools. It is included in this table only for the sake of completeness as most video content today is in the progressive format.

Characteristics and requirements of this application scenario are as follows:

- High encoder complexity (up to 10x and more) can be tolerated since encoding happens once and in parallel for different segments.

- Decoding complexity should be kept at reasonable levels to enable efficient decoder implementation.

- Support and efficient encoding of a wide range of content types and formats is required:

. High Dynamic Range (HDR), Wide Color Gamut (WCG), high resolution (currently, up to 4K), high frame-rate content are important use cases, the codec should be able to encode such content efficiently.

. Coding efficiency improvement at both lower and higher resolutions is important since low resolutions are used when streaming in low bandwidth conditions.

. Improvement on both "easy" and "difficult" content in terms of compression efficiency at the same quality level contributes to the overall bitrate/storage savings.

. Film grain (and sometimes other types of noise) is often present in the streaming movie-type content and is usually a part of the creative intent.

- Significant improvements in compression efficiency between generations of video standards are desirable since this scenario typically assumes long-term support of legacy video codecs.

- Random access points are inserted frequently (one per 2-5 seconds) to enable switching between resolutions and fast-forward playback.

- Elementary stream should have a model that allows easy parsing and identification of the sample components.

- Middle QP values are normally used in streaming, this is also the range where compression efficiency is important for this scenario.

- Scalability or other forms of supporting multiple quality representations are beneficial if they do not incur significant bitrate overhead and if mandated in the first version.

### 3.2. Internet Protocol Television (IPTV)

This is a service for delivering television content over IP-based networks. IPTV may be classified into two main groups based on the type of delivery, as follows:

- unicast (e.g., for video on demand), where delay is not crucial;

- multicast/broadcast (e.g., for transmitting news) where zapping, i.e. stream changing, delay is important.

In the IPTV scenario, traffic is transmitted over managed (QoS-based) networks. Typical content used in this application is news, movies, cartoons, series, TV shows, etc. One important requirement for both groups is Random access to pictures, i.e. random access period (RAP) should be kept small enough (approximately, 1-5 seconds). Optional requirements are as follows:

- Temporal (frame-rate) scalability;

- Resolution and quality (SNR) scalability.

For this application, typical values of resolutions, frame-rates, and RAPs are presented in Table 4.

| Resolution * | Frame-rate, fps | PAM |
|---|---|---|
| 2160p (4K),3840x2160 | 24/1.001, 24, 25, | RA |
| 1080p, 1920x1080 | 30/1.001, 30, 50, | RA |
| 1080i, 1920x1080* | 60/1.001, 60, 100, | RA |
| 720p, 1280x720 | 120/1.001, 120 | RA |
| 576p (EDTV), 720x576 | The set of frame-rates | RA |
| 576i (SDTV), 720x576* | presented in this table | RA |
| 480p (EDTV), 720x480 | is taken from Table 2 | RA |
| 480i (SDTV), 720x480* | in [1] | RA |

Table 4. IPTV: typical values of resolutions, frame-rates, and RAPs

NB *: Interlaced content can be handled at the higher system level and not necessarily by using specialized video coding tools. It is included in this table only for the sake of completeness as most video content today is in progressive format.

### 3.3. Video conferencing

This is a form of video connection over the Internet. This form allows users to establish connections to two or more people by two-way video and audio transmission for communication in real-time. For this application, both stationary and mobile devices can be used. The main requirements are as follows:

- Delay should be kept as low as possible (the preferable and maximum end-to-end delay values should be less than 100 ms [9] and 320 ms [2], respectively);
- Temporal (frame-rate) scalability;
- Error robustness.

Support of resolution and quality (SNR) scalability is highly desirable. For this application, typical values of resolutions, frame-rates, and RAPs are presented in Table 5.

| Resolution | Frame-rate, fps | PAM |
|---|---|---|
| 1080p, 1920x1080 | 15, 30 | FIZD |
| 720p, 1280x720 | 30, 60 | FIZD |
| 4CIF, 704x576 | 30, 60 | FIZD |
| 4SIF, 704x480 | 30, 60 | FIZD |
| VGA, 640x480 | 30, 60 | FIZD |
| 360p, 640x360 | 30, 60 | FIZD |

Table 5. Video conferencing: typical values of resolutions, frame-rates, and RAPs

### 3.4. Video sharing

This is a service that allows people to upload and share video data (using live streaming or not) and to watch them. It is also known as video hosting. A typical User-generated Content (UGC) scenario for this application is to capture video using mobile cameras such as

GoPro or cameras integrated into smartphones (amateur video). The main requirements are as follows:

- Random access to pictures for downloaded video data;
- Temporal (frame-rate) scalability;
- Error robustness.

Support of resolution and quality (SNR) scalability is highly desirable. For this application, typical values of resolutions, frame-rates, and RAPs are presented in Table 6.

| Resolution | Frame-rate, fps | PAM |
|---|---|---|
| 2160p (4K),3840x2160 | 24, 25, 30, 48, 50, 60 | RA |
| 1440p (2K),2560x1440 | 24, 25, 30, 48, 50, 60 | RA |
| 1080p, 1920x1080 | 24, 25, 30, 48, 50, 60 | RA |
| 720p, 1280x720 | 24, 25, 30, 48, 50, 60 | RA |
| 480p, 854x480 | 24, 25, 30, 48, 50, 60 | RA |
| 360p, 640x360 | 24, 25, 30, 48, 50, 60 | RA |

Table 6. Video sharing: typical values of resolutions, frame-rates [10], and RAPs

## 3.5. Screencasting

This is a service that allows users to record and distribute computer desktop screen output. This service requires efficient compression of computer-generated content with high visual quality up to visually and mathematically (numerically) lossless [11]. Currently, this application includes business presentations (powerpoint, word documents, email messages, etc.), animation (cartoons), gaming content, data visualization, i.e. such type of content that is characterized by fast motion, rotation, smooth shade, 3D effect, highly saturated colors with full resolution, clear textures and sharp edges with distinct colors [11]), virtual desktop infrastructure (VDI), screen/desktop sharing and collaboration, supervisory control and data acquisition (SCADA) display, automotive/navigation display, cloud gaming, factory

automation display, wireless display, display wall, digital operating room (DiOR), etc. For this application, an important requirement is the support of low-delay configurations with zero structural delay, a wide range of video formats (e.g., RGB) in addition to YCbCr 4:2:0 and YCbCr 4:4:4 [11]. For this application, typical values of resolutions, frame-rates, and RAPs are presented in Table 7.

| Resolution | Frame-rate, fps | PAM |
|---|---|---|
| Input color format: RGB 4:4:4 | | |
| 5k, 5120x2880 | 15, 30, 60 | AI, RA, FIZD |
| 4k, 3840x2160 | 15, 30, 60 | AI, RA, FIZD |
| WQXGA, 2560x1600 | 15, 30, 60 | AI, RA, FIZD |
| WUXGA, 1920x1200 | 15, 30, 60 | AI, RA, FIZD |
| WSXGA+, 1680x1050 | 15, 30, 60 | AI, RA, FIZD |
| WXGA, 1280x800 | 15, 30, 60 | AI, RA, FIZD |
| XGA, 1024x768 | 15, 30, 60 | AI, RA, FIZD |
| SVGA, 800x600 | 15, 30, 60 | AI, RA, FIZD |
| VGA, 640x480 | 15, 30, 60 | AI, RA, FIZD |
| Input color format: YCbCr 4:4:4 | | |
| 5k, 5120x2880 | 15, 30, 60 | AI, RA, FIZD |
| 4k, 3840x2160 | 15, 30, 60 | AI, RA, FIZD |
| 1440p (2K), 2560x1440 | 15, 30, 60 | AI, RA, FIZD |
| 1080p, 1920x1080 | 15, 30, 60 | AI, RA, FIZD |
| 720p, 1280x720 | 15, 30, 60 | AI, RA, FIZD |

Table 7. Screencasting for RGB and YCbCr 4:4:4 format: typical values of resolutions, frame-rates, and RAPs

### 3.6. Game streaming

This is a service that provides game content over the Internet to different local devices such as notebooks, gaming tablets, etc. In this category of applications, server renders 3D games in cloud server, and streams the game to any device with a wired or wireless broadband connection [12]. There are low latency requirements for transmitting user interactions and receiving game data in less than a turn-around delay of 100 ms. This allows anyone to play (or resume) full featured games from anywhere in the Internet [12]. An example of this application is Nvidia Grid [12]. Another category application is broadcast of video games played by people over the Internet in real time or for later viewing [12]. There are many companies such as Twitch, YY in China enable game broadcasting [12]. Games typically contain a lot of sharp edges and large motion [12]. The main requirements are as follows:

- Random access to pictures for game broadcasting;
- Temporal (frame-rate) scalability;
- Error robustness.

Support of resolution and quality (SNR) scalability is highly desirable. For this application, typical values of resolutions, frame-rates, and RAPs are similar to ones presented in Table 5.

### 3.7. Video monitoring / surveillance

This is a type of live broadcasting over IP-based networks. Video streams are sent to many receivers at the same time. A new receiver may connect to the stream at an arbitrary moment, so random access period should be kept small enough (approximately, ~1-5 seconds). Data are transmitted publicly in the case of video monitoring and privately in the case of video surveillance, respectively. For IP-cameras that have to capture, process and encode video data, complexity including computational and hardware complexity as well as memory bandwidth should be kept low to allow real-time processing. In addition, support of high dynamic range and a monochrome mode (e.g., for infrared cameras) as well as resolution and quality (SNR) scalability is an essential requirement for video surveillance. In some use-cases, high video signal fidelity is required even after lossy compression. Typical values of resolutions, frame-rates, and RAPs for video monitoring / surveillance applications are presented in Table 8.

| Resolution | Frame-rate, fps | PAM |
|---|---|---|
| 2160p (4K),3840x2160 | 12, 25, 30 | RA, FIZD |
| 5Mpixels, 2560x1920 | 12, 25, 30 | RA, FIZD |
| 1080p, 1920x1080 | 25, 30 | RA, FIZD |
| 1.3Mpixels, 1280x960 | 25, 30 | RA, FIZD |
| 720p, 1280x720 | 25, 30 | RA, FIZD |
| SVGA, 800x600 | 25, 30 | RA, FIZD |

Table 8. Video monitoring / surveillance: typical values of resolutions, frame-rates, and RAPs

# 4. Requirements

Taking the requirements discussed above for specific video applications, this chapter proposes requirements for an internet video codec.

## 4.1. General requirements

4.1.1. The most basic requirement is coding efficiency, i.e. compression performance on both "easy" and "difficult" content for applications and use cases in Section 2. The codec should provide higher coding efficiency over state-of-the-art video codecs such as HEVC/H.265 and VP9, at least by 25% in accordance with the methodology described in Section 4.1 of this document. For higher resolutions, the coding efficiency improvements are expected to be higher than for lower resolutions.

4.1.2. Good quality specification and well-defined profiles and levels are required to enable device interoperability and facilitate decoder implementations. A profile consists of a subset of entire bitstream syntax elements and consequently it also defines the necessary tools for decoding a conforming bitstream of that profile. A level imposes a set of numerical limits to the values of some syntax elements. An example of codec levels to be supported is presented in Table 9. An actual level definition should include constraints on features that impact the decoder complexity. For example, these features might be as follows: maximum bit-rate, line buffer size, memory usage, etc.

| Level | Example picture resolution at highest frame rate |
|---|---|
| 1 | 128x96(12,288*)@30.0<br>176x144(25,344*)@15.0 |
| 2 | 352x288(101,376*)@30.0 |
| 3 | 352x288(101,376*)@60.0<br>640x360(230,400*)@30.0 |
| 4 | 640x360(230,400*)@60.0<br>960x540(518,400*)@30.0 |
| 5 | 720x576(414,720*)@75.0<br>960x540(518,400*)@60.0<br>1280x720(921,600*)@30.0 |
| 6 | 1,280x720(921,600*)@68.0<br>2,048x1,080(2,211,840*)@30.0 |
| 7 | 1,280x720(921,600*)@120.0<br>2,048x1,080(2,211,840*)@60.0 |
| 8 | 1,920x1,080(2,073,600*)@120.0<br>3,840x2,160(8,294,400*)@30.0<br>4,096x2,160(8,847,360*)@30.0 |
| 9 | 1,920x1,080(2,073,600*)@250.0<br>4,096x2,160(8,847,360*)@60.0 |
| 10 | 1,920x1,080(2,073,600*)@300.0<br>4,096x2,160(8,847,360*)@120.0 |
| 11 | 3,840x2,160(8,294,400*)@120.0<br>8,192x4,320(35,389,440*)@30.0 |
| 12 | 3,840x2,160(8,294,400*)@250.0<br>8,192x4,320(35,389,440*)@60.0 |
| 13 | 3,840x2,160(8,294,400*)@300.0<br>8,192x4,320(35,389,440*)@120.0 |

Table 9. Codec levels

NB *: The quantities of pixels are presented for such applications where a picture can have an arbitrary size (e.g., screencasting)

4.1.3. Bitstream syntax should allow extensibility and backward compatibility. New features can be supported easily by using metadata (e.g., such as SEI messages, VUI, headers) without affecting the bitstream compatibility with legacy decoders. A newer version of the decoder shall be able to play bitstreams of an older version of the same or lower profile and level.

4.1.4. A bitstream should have a model that allows easy parsing and identification of the sample components (such as ISO/IEC14496-10, Annex B or ISO/IEC 14496-15). In particular, information needed for packet handling (e.g., frame type) should not require parsing anything below the header level.

4.1.5. Perceptual quality tools (such as adaptive QP and quantization matrices) should be supported by the codec bit-stream.

4.1.6. The codec specification shall define a buffer model such as hypothetical reference decoder (HRD).

4.1.7. Specifications providing integration with system and delivery layers should be developed.

4.2. Basic requirements

4.2.1. Input source formats:

- Bit depth: 8- and 10-bits (up to 12-bits for a high profile) per color component;
- Color sampling formats:
  - YCbCr 4:2:0;
  - YCbCr 4:4:4, YCbCr 4:2:2 and YCbCr 4:0:0 (preferably in different profile(s)).
- For profiles with bit depth of 10 bits per sample or higher, support of high dynamic range and wide color gamut.
- Support of arbitrary resolution according to the level constraints for such applications where a picture can have an arbitrary size (e.g., in screencasting).
- Exemplary input source formats for codec profiles are shown in Table 10.

| Profile | Bit-depths per color component | Color sampling formats |
|---|---|---|
| 1 | 8 and 10 | 4:0:0 and 4:2:0 |
| 2 | 8 and 10 | 4:0:0, 4:2:0 and 4:4:4 |
| 3 | 8, 10 and 12 | 4:0:0, 4:2:0, 4:2:2 and 4:4:4 |

Table 10. Exemplary input source formats for codec profiles

4.2.2. Coding delay:

- Support of configurations with zero structural delay also referred to as "low-delay" configurations.

  . Note 1: end-to-end delay should be up to 320 ms [2] but its preferable value should be less than 100 ms [9]

- Support of efficient random access point encoding (such as intra coding and resetting of context variables) as well as efficient switching between multiple quality representations.

- Support of configurations with non-zero structural delay (such as out-of-order or multi-pass encoding) for applications without low-delay requirements if such configurations provide additional compression efficiency improvements.

4.2.3. Complexity:

- Feasible real-time implementation of both an encoder and a decoder supporting a chosen subset of tools for hardware and software implementation on a wide range of state-of-the-art platforms. The real-time encoder tools subset should provide meaningful improvement in compression efficiency at reasonable complexity of hardware and software encoder implementations as compared to real-time implementations of state-of-the-art video compression technologies such as HEVC/H.265 and VP9.

- High-complexity software encoder implementations used by offline encoding applications can have 10x or more complexity increase compared to state-of-the-art video compression technologies such as HEVC/H.265 and VP9.

4.2.4. Scalability:

- Temporal (frame-rate) scalability should be supported.

4.2.5. Error resilience:

- Error resilience tools that are complementary to the error protection mechanisms implemented on transport level should be supported.
- The codec should support mechanisms that facilitate packetization of a bitstream for common network protocols.
- Packetization mechanisms should enable frame-level error recovery by means of retransmission or error concealment.
- The codec should support effective mechanisms for allowing decoding and reconstruction of significant parts of pictures in the event that parts of the picture data are lost in transmission.
- The bitstream specification shall support independently decodable sub-frame units similar to slices or independent tiles. It shall be possible for the encoder to restrict the bit-stream to allow parsing of the bit-stream after a packet-loss and to communicate it to the decoder.

4.3. Optional requirements

4.3.1. Input source formats

- Bit depth: up to 16-bits per color component.
- Color sampling formats: RGB 4:4:4.
- Auxiliary channel (e.g., alpha channel) support.

4.3.2. Scalability:

- Resolution and quality (SNR) scalability that provide low compression efficiency penalty (up to 5% of BD-rate [13] increase per layer with reasonable increase of both computational and hardware complexity) can be supported in the main profile of the codec being developed by the NETVC WG. Otherwise, a separate profile is needed to support these types of scalability.

- Computational complexity scalability(i.e. computational complexity is decreasing along with degrading picture quality) is desirable.

4.3.3. Complexity:

Tools that enable parallel processing (e.g., slices, tiles, wave front propagation processing) at both encoder and decoder sides are highly desirable for many applications.

- High-level multi-core parallelism: encoder and decoder operation, especially entropy encoding and decoding, should allow multiple frames or sub-frame regions (e.g. 1D slices, 2D tiles, or partitions) to be processed concurrently, either independently or with deterministic dependencies that can be efficiently pipelined

- Low-level instruction set parallelism: favor algorithms that are SIMD/GPU friendly over inherently serial algorithms

4.3.4. Coding efficiency

Compression efficiency on noisy content, content with film grain, computer generated content, and low resolution materials is desirable.

## 5. Evaluation methodology

As shown in Fig.1, compression performance testing is performed in 3 overlapped ranges that encompass 10 different bitrate values:

- Low bitrate range (LBR) is the range that contains the 4 lowest bitrates of the 10 specified bitrates (1 of the 4 bitrate values is shared with the neighboring range);

- Medium bitrate range (MBR) is the range that contains the 4 medium bitrates of the 10 specified bitrates (2 of the 4 bitrate values are shared with the neighboring ranges);

- High bitrate range (HBR) is the range that contains the 4 highest bitrates of the 10 specified bitrates (1 of the 4 bitrate values is shared with the neighboring range).

Initially, for the codec selected as a reference one (e.g., HEVC or VP9), a set of 10 QP (quantization parameter) values should be specified in [14] and corresponding quality values should be calculated. In Fig.1, QP and quality values are denoted as QP0, QP1, QP2,..., QP8, QP9 and Q0, Q1, Q2,..., Q8, Q9, respectively. To

guarantee the overlaps of quality levels between the bitrate ranges of the reference and tested codecs, a quality alignment procedure should be performed for each range's outermost (left- and rightmost) quality levels Qk of the reference codec (i.e. for Q0, Q3, Q6, and Q9) and the quality levels Q'k (i.e. Q'0, Q'3, Q'6, and Q'9) of the tested codec. Thus, these quality levels Q'k and, hence, the corresponding QP value QP'k (i.e. QP'0, QP'3, QP'6, and QP'9) of the tested codec should be selected using the following formulas:

$$Q'k = \min_{i \in R} \{ abs(Q'i - Qk) \},$$

$$QP'k = \operatorname{argmin}_{i \in R} \{ abs(Q'i(QP'i) - Qk(QPk)) \},$$

where R is the range of the QP indexes of the tested codec, i.e. the candidate Internet video codec. The inner quality levels (i.e. Q'1, Q'2, Q'4, Q'5, Q'7, and Q'8) as well as their corresponding QP values of each range (i.e. QP'1, QP'2, QP'4, QP'5, QP'7, and QP'8) should be as equidistantly spaced as possible between the left- and rightmost quality levels without explicitly mapping their values using the above described procedure.

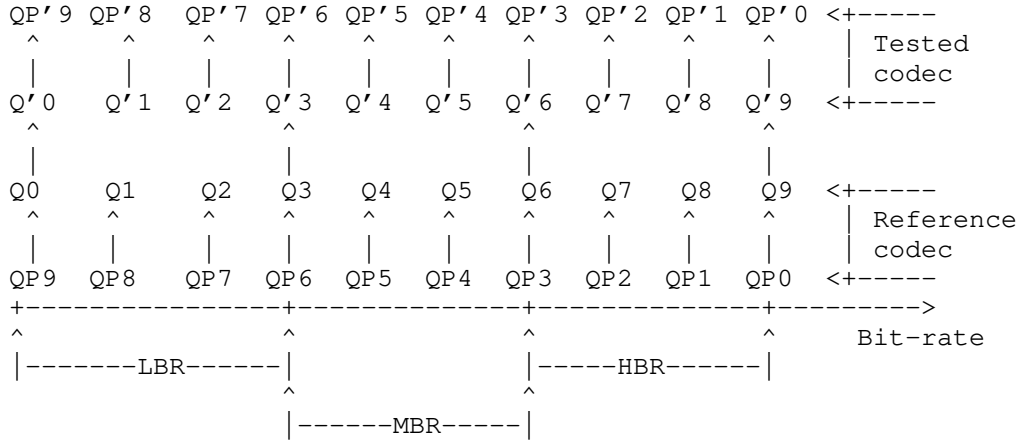



Figure 1 Quality/QP alignment for compression performance evaluation

Since the QP mapping results may vary for different sequences, eventually, this quality alignment procedure needs to be separately performed for each quality assessment index and each sequence used for codec performance evaluation to fulfill the above described requirements.

To assess the quality of output (decoded) sequences, two indexes, PSNR [3] and MS-SSIM [3,15] are separately computed. In the case of the YCbCr color format, PSNR should be calculated for each color plane whereas MS-SSIM is calculated for luma channel only. In the case of the RGB color format, both metrics are computed for R, G and B channels. Thus, for each sequence, 30 RD-points for PSNR (i.e. three RD-curves, one for each channel) and 10 RD-points for MS-SSIM (i.e. one RD-curve, for luma channel only) should be calculated in the case of YCbCr. If content is encoded as RGB, 60 RD-points (30 for PSNR and 30 for MS-SSIM) should be calculated, i.e. three RD-curves (one for each channel) are computed for PSNR as well as three RD-curves (one for each channel) for MS-SSIM.

Finally, to obtain an integral estimation, BD-rate savings [13] should be computed for each range and each quality index. In addition, average values over all the 3 ranges should be provided for both PSNR and MS-SSIM. A list of video sequences that should be used for testing as well as the 10 QP values for the reference codec are defined in [14]. Testing processes should use the information on the codec applications presented in this document. As the reference for evaluation, state-of-the-art video codecs such as HEVC/H.265 [4,5] or VP9 must be used. The reference source code of the HEVC/H.265 codec can be found at [6]. The HEVC/H.265 codec must be configured according to [16] and Table 11.

| Intra-period, second | HEVC/H.265 encoding mode according to [16] |
|---|---|
| AI | Intra Main or Intra Main10 |
| RA | Random access Main or Random access Main10 |
| FIZD | Low delay Main or Low delay Main10 |

Table 11. Intra-periods for different HEVC/H.265 encoding modes according to [16]

According to the coding efficiency requirement described in Section 3.1.1, BD-rate savings calculated for each color plane and averaged for all the video sequences used to test the NETVC codec should be, at least,

o 25% if calculated over the whole bitrate range;

o 15% if calculated for each bitrate subrange (LBR, MBR, HBR).

Since values of the two objective metrics (PSNR and MS-SSIM) are available for some color planes, each value should meet these coding efficiency requirements, i.e. the final BD-rate saving denoted as S is calculated for a given color plane as follows:

S = min { S_psnr, S_ms-ssim },

where S_psnr and S_ms-ssim are BD-rate savings calculated for the given color plane using PSNR and MS-SSIM metrics, respectively.

In addition to the objective quality measures defined above, subjective evaluation must also be performed for the final NETVC codec adoption. For subjective tests, the MOS-based evaluation procedure must be used as described in section 2.1 of [3]. For perception-oriented tools that primarily impact subjective quality, additional tests may also be individually assigned even for intermediate evaluation, subject to a decision of the NETVC WG.

## 6. Security Considerations

This document itself does not address any security considerations. However, it is worth noting that a codec implementation (for both an encoder and a decoder) should take into consideration the worst-case computational complexity, memory bandwidth, and physical memory size needed to processes the potentially untrusted input (e.g., the decoded pictures used as references).

## 7. IANA Considerations

This document has no IANA actions.

## 8. References

### 8.1. Normative References

[1] Recommendation ITU-R BT.2020-2: Parameter values for ultra-high definition television systems for production and international programme exchange, 2015.

[2] Recommendation ITU-T G.1091: Quality of Experience requirements for telepresence services, 2014.

[3] ISO/IEC PDTR 29170-1: Information technology -- Advanced image coding and evaluation methodologies -- Part 1: Guidelines for codec evaluation.

[4] ISO/IEC 23008-2:2015. Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Part 2: High efficiency video coding

[5] Recommendation ITU-T H.265: High efficiency video coding, 2013.

[6] High Efficiency Video Coding (HEVC) reference software (HEVC Test Model also known as HM) at the web-site of Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute (HHI): https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/

## 8.2. Informative References

[7] Definition of the term "high dynamic range imaging" at the web-site of Federal Agencies Digital Guidelines Initiative: http://www.digitizationguidelines.gov/term.php?term=highdynamicrangeimaging

[8] Definition of the term "compression, visually lossless" at the web-site of Federal Agencies Digital Guidelines Initiative: http://www.digitizationguidelines.gov/term.php?term=compressionvisuallylossless

[9] S. Wenger, "The case for scalability support in version 1 of Future Video Coding," Document COM 16-C 988 R1-E of ITU-T Video Coding Experts Group (ITU-T Q.6/SG 16), Geneva, Switzerland, September 2015.

[10] "Recommended upload encoding settings (Advanced)" for the YouTube video-sharing service: https://support.google.com/youtube/answer/1722171?hl=en

[11] H. Yu, K. McCann, R. Cohen, and P. Amon, "Requirements for future extensions of HEVC in coding screen content", Document N14174 of Moving Picture Experts Group (ISO/IEC JTC 1/SC 29/ WG 11), San Jose, USA, January 2014.

[12] Manindra Parhy, "Game streaming requirement for Future Video Coding," Document N36771 of ISO/IEC Moving Picture Experts Group (ISO/IEC JTC 1/SC 29/WG 11), Warsaw, Poland, June 2015.

[13] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," Document VCEG-M33 of ITU-T Video Coding Experts Group (ITU-T Q.6/SG 16), Austin, Texas, USA, April 2001.

[14] T. Daede, A. Norkin, and I. Brailovskiy, "Video Codec Testing and Quality Measurement", draft-ietf-netvc-testing-08(work in progress), January 2019, p.23.

[15] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," Invited Paper, IEEE Asilomar Conference on Signals, Systems and Computers, Nov. 2003, Vol. 2, pp. 1398-1402.

[16] F. Bossen, "Common test conditions and software reference configurations," Document JCTVC-L1100 of Joint Collaborative Team on Video Coding (JCT-VC) of the ITU-T Video Coding Experts Group (ITU-T Q.6/SG 16) and ISO/IEC Moving Picture Experts Group (ISO/IEC JTC 1/SC 29/WG 11), Geneva, Switzerland, January 2013.

## 9. Acknowledgments

The authors would like to thank Mr. Paul Coverdale, Mr. Vasily Rufitskiy, and Dr. Jianle Chen for many useful discussions on this document and their help while preparing it as well as Mr. Mo Zanaty, Dr. Minhua Zhou, Dr. Ali Begen, Mr. Thomas Daede, Mr. Adam Roach, Dr. Thomas Davies, Mr. Jonathan Lennox, Dr. Timothy Terriberry, Mr. Peter Thatcher, Dr. Jean-Marc Valin, Mr. Roman Danyliw, Mr. Jack Moffitt, Mr. Greg Coppa, and Mr. Andrew Krupiczka for their valuable comments on different revisions of this document.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Alexey Filippov
Huawei Technologies

Email: alexey.filippov@huawei.com

Andrey Norkin
Netflix

Email: anorkin@netflix.com

Jose Roberto Alvarez
Huawei Technologies

Email: j.alvarez@ieee.org

Network Working Group
Internet-Draft
Intended status: Informational
Expires: August 3, 2020

T. Daede
Mozilla
A. Norkin
Netflix
I. Brailovskiy
Amazon Lab126

January 31, 2020

# Video Codec Testing and Quality Measurement
draft-ietf-netvc-testing-09

## Abstract

This document describes guidelines and procedures for evaluating a video codec. This covers subjective and objective tests, test conditions, and materials used for the test.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 3, 2020.

## Copyright Notice

Table of Contents

## 1. Introduction

When developing a video codec, changes and additions to the codec need to be decided based on their performance tradeoffs. In addition, measurements are needed to determine when the codec has met its performance goals. This document specifies how the tests are to be carried about to ensure valid comparisons when evaluating changes under consideration. Authors of features or changes should provide the results of the appropriate test when proposing codec modifications.

## 2. Subjective quality tests

Subjective testing uses human viewers to rate and compare the quality of videos. It is the preferable method of testing video codecs.

Subjective testing results take priority over objective testing results, when available. Subjective testing is recommended especially when taking advantage of psychovisual effects that may not be well represented by objective metrics, or when different objective metrics disagree.

Selection of a testing methodology depends on the feature being tested and the resources available. Test methodologies are presented in order of increasing accuracy and cost.

Testing relies on the resources of participants. If a participant requires a subjective test for a particular feature or improvement, they are responsible for ensuring that resources are available. This ensures that only important tests be done; in particular, the tests that are important to participants.

Subjective tests should use the same operating points as the objective tests.

### 2.1. Still Image Pair Comparison

A simple way to determine superiority of one compressed image is to visually compare two compressed images, and have the viewer judge which one has a higher quality. For example, this test may be suitable for an intra de-ringing filter, but not for a new inter prediction mode. For this test, the two compressed images should have similar compressed file sizes, with one image being no more than 5% larger than the other. In addition, at least 5 different images should be compared.

Once testing is complete, a p-value can be computed using the binomial test. A significant result should have a resulting p-value less than or equal to 0.5. For example:

p_value = binom_test(a,a+b)

where a is the number of votes for one video, b is the number of votes for the second video, and binom_test(x,y) returns the binomial PMF (probability mass function) with x observed tests, y total tests, and expected probability 0.5.

If ties are allowed to be reported, then the equation is modified:

p_value = binom_test(a+floor(t/2),a+b+t)

where t is the number of tie votes.

Still image pair comparison is used for rapid comparisons during development - the viewer may be either a developer or user, for example. As the results are only relative, it is effective even with an inconsistent viewing environment. Because this test only uses still images (keyframes), this is only suitable for changes with similar or no effect on inter frames.

## 2.2. Video Pair Comparison

The still image pair comparison method can be modified to also compare vidoes. This is necessary when making changes with temporal effects, such as changes to inter-frame prediction. Video pair comparisons follow the same procedure as still images. Videos used for testing should be limited to 10 seconds in length, and can be rewatched an unlimited number of times.

## 2.3. Mean Opinion Score

A Mean Opinion Score (MOS) viewing test is the preferred method of evaluating the quality. The subjective test should be performed as either consecutively showing the video sequences on one screen or on two screens located side-by-side. The testing procedure should normally follow rules described in [BT500] and be performed with non-expert test subjects. The result of the test will be (depending on the test procedure) mean opinion scores (MOS) or differential mean opinion scores (DMOS). Confidence intervals are also calculated to judge whether the difference between two encodings is statistically significant. In certain cases, a viewing test with expert test subjects can be performed, for example if a test should evaluate technologies with similar performance with respect to a particular artifact (e.g. loop filters or motion prediction). Unlike pair

comparisions, a MOS test requires a consistent testing environment. This means that for large scale or distributed tests, pair comparisons are preferred.

## 3. Objective Metrics

Objective metrics are used in place of subjective metrics for easy and repeatable experiments. Most objective metrics have been designed to correlate with subjective scores.

The following descriptions give an overview of the operation of each of the metrics. Because implementation details can sometimes vary, the exact implementation is specified in C in the Daala tools repository [DAALA-GIT]. Implementations of metrics must directly support the input's resolution, bit depth, and sampling format.

Unless otherwise specified, all of the metrics described below only apply to the luma plane, individually by frame. When applied to the video, the scores of each frame are averaged to create the final score.

Codecs must output the same resolution, bit depth, and sampling format as the input.

### 3.1. Overall PSNR

PSNR is a traditional signal quality metric, measured in decibels. It is directly drived from mean square error (MSE), or its square root (RMSE). The formula used is:

20 * log10 ( MAX / RMSE )

or, equivalently:

10 * log10 ( MAX^2 / MSE )

where the error is computed over all the pixels in the video, which is the method used in the dump_psnr.c reference implementation.

This metric may be applied to both the luma and chroma planes, with all planes reported separately.

### 3.2. Frame-averaged PSNR

PSNR can also be calculated per-frame, and then the values averaged together. This is reported in the same way as overall PSNR.

### 3.3. PSNR-HVS-M

The PSNR-HVS [PSNRHVS] metric performs a DCT transform of 8x8 blocks of the image, weights the coefficients, and then calculates the PSNR of those coefficients. Several different sets of weights have been considered. The weights used by the dump_pnsrhvs.c tool in the Daala repository have been found to be the best match to real MOS scores.

### 3.4. SSIM

SSIM (Structural Similarity Image Metric) is a still image quality metric introduced in 2004 [SSIM]. It computes a score for each individual pixel, using a window of neighboring pixels. These scores can then be averaged to produce a global score for the entire image. The original paper produces scores ranging between 0 and 1.

To linearize the metric for BD-Rate computation, the score is converted into a nonlinear decibel scale:

-10 * log10 (1 - SSIM)

### 3.5. Multi-Scale SSIM

Multi-Scale SSIM is SSIM extended to multiple window sizes [MSSSIM]. The metric score is converted to decibels in the same way as SSIM.

### 3.6. CIEDE2000

CIEDE2000 is a metric based on CIEDE color distances [CIEDE2000]. It generates a single score taking into account all three chroma planes. It does not take into consideration any structural similarity or other psychovisual effects.

### 3.7. VMAF

Video Multi-method Assessment Fusion (VMAF) is a full-reference perceptual video quality metric that aims to approximate human perception of video quality [VMAF]. This metric is focused on quality degradation due to compression and rescaling. VMAF estimates the perceived quality score by computing scores from multiple quality assessment algorithms, and fusing them using a support vector machine (SVM). Currently, three image fidelity metrics and one temporal signal have been chosen as features to the SVM, namely Anti-noise SNR (ANSNR), Detail Loss Measure (DLM), Visual Information Fidelity (VIF), and the mean co-located pixel difference of a frame with respect to the previous frame.

The quality score from VMAF is used directly to calculate BD-Rate, without any conversions.

# 4. Comparing and Interpreting Results

## 4.1. Graphing

When displayed on a graph, bitrate is shown on the X axis, and the quality metric is on the Y axis. For publication, the X axis should be linear. The Y axis metric should be plotted in decibels. If the quality metric does not natively report quality in decibels, it should be converted as described in the previous section.

## 4.2. BD-Rate

The Bjontegaard rate difference, also known as BD-rate, allows the measurement of the bitrate reduction offered by a codec or codec feature, while maintaining the same quality as measured by objective metrics. The rate change is computed as the average percent difference in rate over a range of qualities. Metric score ranges are not static - they are calculated either from a range of bitrates of the reference codec, or from quantizers of a third, anchor codec. Given a reference codec and test codec, BD-rate values are calculated as follows:

- Rate/distortion points are calculated for the reference and test codec.
  - At least four points must be computed. These points should be the same quantizers when comparing two versions of the same codec.
  - Additional points outside of the range should be discarded.
- The rates are converted into log-rates.
- A piecewise cubic hermite interpolating polynomial is fit to the points for each codec to produce functions of log-rate in terms of distortion.
- Metric score ranges are computed:
  - If comparing two versions of the same codec, the overlap is the intersection of the two curves, bound by the chosen quantizer points.
  - If comparing dissimilar codecs, a third anchor codec's metric scores at fixed quantizers are used directly as the bounds.

- The log-rate is numerically integrated over the metric range for each curve, using at least 1000 samples and trapezoidal integration.

- The resulting integrated log-rates are converted back into linear rate, and then the percent difference is calculated from the reference to the test codec.

### 4.3. Ranges

For individual feature changes in libaom or libvpx, the overlap BD-Rate method with quantizers 20, 32, 43, and 55 must be used.

For the final evaluation described in [I-D.ietf-netvc-requirements], the quantizers used are 20, 24, 28, 32, 36, 39, 43, 47, 51, and 55.

## 5. Test Sequences

### 5.1. Sources

Lossless test clips are preferred for most tests, because the structure of compression artifacts in already-compressed clips may introduce extra noise in the test results. However, a large amount of content on the internet needs to be recompressed at least once, so some sources of this nature are useful. The encoder should run at the same bit depth as the original source. In addition, metrics need to support operation at high bit depth. If one or more codecs in a comparison do not support high bit depth, sources need to be converted once before entering the encoder.

### 5.2. Test Sets

Sources are divided into several categories to test different scenarios the codec will be required to operate in. For easier comparison, all videos in each set should have the same color subsampling, same resolution, and same number of frames. In addition, all test videos must be publicly available for testing use, to allow for reproducibility of results. All current test sets are available for download [TESTSEQUENCES].

Test sequences should be downloaded in whole. They should not be recreated from the original sources.

Each clip is labeled with its resolution, bit depth, color subsampling, and length.

#### 5.2.1. regression-1

This test set is used for basic regression testing. It contains a very small number of clips.

- kirlandvga (640x360, 8bit, 4:2:0, 300 frames)
- FourPeople (1280x720, 8bit, 4:2:0, 60 frames)
- Narrarator (4096x2160, 10bit, 4:2:0, 15 frames)
- CSGO (1920x1080, 8bit, 4:4:4 60 frames)

#### 5.2.2. objective-2-slow

This test set is a comprehensive test set, grouped by resolution. These test clips were created from originals at [TESTSEQUENCES]. They have been scaled and cropped to match the resolution of their category. This test set requires a codec that supports both 8 and 10 bit video.

4096x2160, 4:2:0, 60 frames:

- Netflix_BarScene_4096x2160_60fps_10bit_420_60f
- Netflix_BoxingPractice_4096x2160_60fps_10bit_420_60f
- Netflix_Dancers_4096x2160_60fps_10bit_420_60f
- Netflix_Narrator_4096x2160_60fps_10bit_420_60f
- Netflix_RitualDance_4096x2160_60fps_10bit_420_60f
- Netflix_ToddlerFountain_4096x2160_60fps_10bit_420_60f
- Netflix_WindAndNature_4096x2160_60fps_10bit_420_60f
- street_hdr_amazon_2160p

1920x1080, 4:2:0, 60 frames:

- aspen_1080p_60f
- crowd_run_1080p50_60f
- ducks_take_off_1080p50_60f
- guitar_hdr_amazon_1080p

- life_1080p30_60f
- Netflix_Aerial_1920x1080_60fps_8bit_420_60f
- Netflix_Boat_1920x1080_60fps_8bit_420_60f
- Netflix_Crosswalk_1920x1080_60fps_8bit_420_60f
- Netflix_FoodMarket_1920x1080_60fps_8bit_420_60f
- Netflix_PierSeaside_1920x1080_60fps_8bit_420_60f
- Netflix_SquareAndTimelapse_1920x1080_60fps_8bit_420_60f
- Netflix_TunnelFlag_1920x1080_60fps_8bit_420_60f
- old_town_cross_1080p50_60f
- pan_hdr_amazon_1080p
- park_joy_1080p50_60f
- pedestrian_area_1080p25_60f
- rush_field_cuts_1080p_60f
- rush_hour_1080p25_60f
- seaplane_hdr_amazon_1080p
- station2_1080p25_60f
- touchdown_pass_1080p_60f

1280x720, 4:2:0, 120 frames:

- boat_hdr_amazon_720p
- dark720p_120f
- FourPeople_1280x720_60_120f
- gipsrestat720p_120f
- Johnny_1280x720_60_120f
- KristenAndSara_1280x720_60_120f

- Netflix_DinnerScene_1280x720_60fps_8bit_420_120f
- Netflix_DrivingPOV_1280x720_60fps_8bit_420_120f
- Netflix_FoodMarket2_1280x720_60fps_8bit_420_120f
- Netflix_RollerCoaster_1280x720_60fps_8bit_420_120f
- Netflix_Tango_1280x720_60fps_8bit_420_120f
- rain_hdr_amazon_720p
- vidyo1_720p_60fps_120f
- vidyo3_720p_60fps_120f
- vidyo4_720p_60fps_120f

640x360, 4:2:0, 120 frames:

- blue_sky_360p_120f
- controlled_burn_640x360_120f
- desktop2360p_120f
- kirland360p_120f
- mmstationary360p_120f
- niklas360p_120f
- rain2_hdr_amazon_360p
- red_kayak_360p_120f
- riverbed_360p25_120f
- shields2_640x360_120f
- snow_mnt_640x360_120f
- speed_bag_640x360_120f
- stockholm_640x360_120f
- tacomanarrows360p_120f

- thaloundeskmtg360p_120f
- water_hdr_amazon_360p

426x240, 4:2:0, 120 frames:

- bqfree_240p_120f
- bqhighway_240p_120f
- bqzoom_240p_120f
- chairlift_240p_120f
- dirtbike_240p_120f
- mozzoom_240p_120f

1920x1080, 4:4:4 or 4:2:0, 60 frames:

- CSGO_60f.y4m
- DOTA2_60f_420.y4m
- MINECRAFT_60f_420.y4m
- STARCRAFT_60f_420.y4m
- EuroTruckSimulator2_60f.y4m
- Hearthstone_60f.y4m
- wikipedia_420.y4m
- pvq_slideshow.y4m

#### 5.2.3. objective-2-fast

This test set is a strict subset of objective-2-slow. It is designed for faster runtime. This test set requires compiling with high bit depth support.

1920x1080, 4:2:0, 60 frames:

- aspen_1080p_60f
- ducks_take_off_1080p50_60f

- life_1080p30_60f
- Netflix_Aerial_1920x1080_60fps_8bit_420_60f
- Netflix_Boat_1920x1080_60fps_8bit_420_60f
- Netflix_FoodMarket_1920x1080_60fps_8bit_420_60f
- Netflix_PierSeaside_1920x1080_60fps_8bit_420_60f
- Netflix_SquareAndTimelapse_1920x1080_60fps_8bit_420_60f
- Netflix_TunnelFlag_1920x1080_60fps_8bit_420_60f
- rush_hour_1080p25_60f
- seaplane_hdr_amazon_1080p
- touchdown_pass_1080p_60f

1280x720, 4:2:0, 120 frames:

- boat_hdr_amazon_720p
- dark720p_120f
- gipsrestat720p_120f
- KristenAndSara_1280x720_60_120f
- Netflix_DrivingPOV_1280x720_60fps_8bit_420_60f
- Netflix_RollerCoaster_1280x720_60fps_8bit_420_60f
- vidyo1_720p_60fps_120f
- vidyo4_720p_60fps_120f

640x360, 4:2:0, 120 frames:

- blue_sky_360p_120f
- controlled_burn_640x360_120f
- kirland360p_120f
- niklas360p_120f

- rain2_hdr_amazon_360p
- red_kayak_360p_120f
- riverbed_360p25_120f
- shields2_640x360_120f
- speed_bag_640x360_120f
- thaloundeskmtg360p_120f

426x240, 4:2:0, 120 frames:

- bqfree_240p_120f
- bqzoom_240p_120f
- dirtbike_240p_120f

1290x1080, 4:2:0, 60 frames:

- DOTA2_60f_420.y4m
- MINECRAFT_60f_420.y4m
- STARCRAFT_60f_420.y4m
- wikipedia_420.y4m

#### 5.2.4. objective-1.1

This test set is an old version of objective-2-slow.

4096x2160, 10bit, 4:2:0, 60 frames:

- Aerial (start frame 600)
- BarScene (start frame 120)
- Boat (start frame 0)
- BoxingPractice (start frame 0)
- Crosswalk (start frame 0)
- Dancers (start frame 120)

- FoodMarket
- Narrator
- PierSeaside
- RitualDance
- SquareAndTimelapse
- ToddlerFountain (start frame 120)
- TunnelFlag
- WindAndNature (start frame 120)

1920x1080, 8bit, 4:4:4, 60 frames:

- CSGO
- DOTA2
- EuroTruckSimulator2
- Hearthstone
- MINECRAFT
- STARCRAFT
- wikipedia
- pvq_slideshow

1920x1080, 8bit, 4:2:0, 60 frames:

- ducks_take_off
- life
- aspen
- crowd_run
- old_town_cross
- park_joy

- pedestrian_area
- rush_field_cuts
- rush_hour
- station2
- touchdown_pass

1280x720, 8bit, 4:2:0, 60 frames:

- Netflix_FoodMarket2
- Netflix_Tango
- DrivingPOV (start frame 120)
- DinnerScene (start frame 120)
- RollerCoaster (start frame 600)
- FourPeople
- Johnny
- KristenAndSara
- vidyo1
- vidyo3
- vidyo4
- dark720p
- gipsrecmotion720p
- gipsrestat720p
- controlled_burn
- stockholm
- speed_bag
- snow_mnt

- shields

640x360, 8bit, 4:2:0, 60 frames:

- red_kayak
- blue_sky
- riverbed
- thaloundeskmtgvga
- kirlandvga
- tacomanarrowsvga
- tacomascmvvga
- desktop2360p
- mmmovingvga
- mmstationaryvga
- niklasvga

#### 5.2.5. objective-1-fast

This is an old version of objective-2-fast.

1920x1080, 8bit, 4:2:0, 60 frames:

- Aerial (start frame 600)
- Boat (start frame 0)
- Crosswalk (start frame 0)
- FoodMarket
- PierSeaside
- SquareAndTimelapse
- TunnelFlag

1920x1080, 8bit, 4:2:0, 60 frames:

- CSGO
- EuroTruckSimulator2
- MINECRAFT
- wikipedia

1920x1080, 8bit, 4:2:0, 60 frames:

- ducks_take_off
- aspen
- old_town_cross
- pedestrian_area
- rush_hour
- touchdown_pass

1280x720, 8bit, 4:2:0, 60 frames:

- Netflix_FoodMarket2
- DrivingPOV (start frame 120)
- RollerCoaster (start frame 600)
- Johnny
- vidyo1
- vidyo4
- gipsrecmotion720p
- speed_bag
- shields

640x360, 8bit, 4:2:0, 60 frames:

- red_kayak
- riverbed

- kirlandvga
- tacomascmvvga
- mmmovingvga
- niklasvga

### 5.3. Operating Points

Four operating modes are defined. High latency is intended for on demand streaming, one-to-many live streaming, and stored video. Low latency is intended for videoconferencing and remote access. Both of these modes come in CQP (constant quantizer parameter) and unconstrained variants. When testing still image sets, such as subset1, high latency CQP mode should be used.

#### 5.3.1. Common settings

Encoders should be configured to their best settings when being compared against each other:

- av1: -codec=av1 -ivf -frame-parallel=0 -tile-columns=0 -cpu-used=0 -threads=1

#### 5.3.2. High Latency CQP

High Latency CQP is used for evaluating incremental changes to a codec. This method is well suited to compare codecs with similar coding tools. It allows codec features with intrinsic frame delay.

- daala: -v=x -b 2
- vp9: -end-usage=q -cq-level=x -lag-in-frames=25 -auto-alt-ref=2
- av1: -end-usage=q -cq-level=x -auto-alt-ref=2

#### 5.3.3. Low Latency CQP

Low Latency CQP is used for evaluating incremental changes to a codec. This method is well suited to compare codecs with similar coding tools. It requires the codec to be set for zero intrinsic frame delay.

- daala: -v=x
- av1: -end-usage=q -cq-level=x -lag-in-frames=0

#### 5.3.4. Unconstrained High Latency

The encoder should be run at the best quality mode available, using the mode that will provide the best quality per bitrate (VBR or constant quality mode). Lookahead and/or two-pass are allowed, if supported. One parameter is provided to adjust bitrate, but the units are arbitrary. Example configurations follow:

- x264: --crf=x
- x265: --crf=x
- daala: -v=x -b 2
- av1: --end-usage=q --cq-level=x --lag-in-frames=25 --auto-alt-ref=2

#### 5.3.5. Unconstrained Low Latency

The encoder should be run at the best quality mode available, using the mode that will provide the best quality per bitrate (VBR or constant quality mode), but no frame delay, buffering, or lookahead is allowed. One parameter is provided to adjust bitrate, but the units are arbitrary. Example configurations follow:

- x264: --crf-x --tune zerolatency
- x265: --crf=x --tune zerolatency
- daala: -v=x
- av1: --end-usage=q --cq-level=x --lag-in-frames=0

## 6. Automation

Frequent objective comparisons are extremely beneficial while developing a new codec. Several tools exist in order to automate the process of objective comparisons. The Compare-Codecs tool allows BD-rate curves to be generated for a wide variety of codecs [COMPARECODECS]. The Daala source repository contains a set of scripts that can be used to automate the various metrics used. In addition, these scripts can be run automatically utilizing distributed computers for fast results, with rd_tool [RD_TOOL]. This tool can be run via a web interface called AreWeCompressedYet [AWCY], or locally.

Because of computational constraints, several levels of testing are specified.

### 6.1. Regression tests

Regression tests run on a small number of short sequences - regression-test-1. The regression tests should include a number of various test conditions. The purpose of regression tests is to ensure bug fixes (and similar patches) do not negatively affect the performance. The anchor in regression tests is the previous revision of the codec in source control. Regression tests are run on both high and low latency CQP modes

### 6.2. Objective performance tests

Changes that are expected to affect the quality of encode or bitstream should run an objective performance test. The performance tests should be run on a wider number of sequences. The following data should be reported:

- Identifying information for the encoder used, such as the git commit hash.
- Command line options to the encoder, configure script, and anything else necessary to replicate the experiment.
- The name of the test set run (objective-1-fast)
- For both high and low latency CQP modes, and for each objective metric:
  - The BD-Rate score, in percent, for each clip.
  - The average of all BD-Rate scores, equally weighted, for each resolution category in the test set.
  - The average of all BD-Rate scores for all videos in all categories.

Normally, the encoder should always be run at the slowest, highest quality speed setting (cpu-used=0 in the case of AV1 and VP9). However, in the case of computation time, both the reference and changed encoder can be built with some options disabled. For AV1, -disable-ext_partition and -disable-ext_partition_types can be passed to the configure script to substantially speed up encoding, but the usage of these options must be reported in the test results.

### 6.3. Periodic tests

Periodic tests are run on a wide range of bitrates in order to gauge progress over time, as well as detect potential regressions missed by other tests.

## 7. IANA Considerations

This document does not require any IANA actions.

## 8. Security Considerations

This document describes the methodologies an procedures for qualitative testing, therefore does not iteself have implications for network of decoder security.

## 9. Informative References

[AWCY] Xiph.Org, "Are We Compressed Yet?", 2016, <https://arewecompressedyet.com/>.

[BT500] ITU-R, "Recommendation ITU-R BT.500-13", 2012, <https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf>.

[CIEDE2000] Yang, Y., Ming, J., and N. Yu, "Color Image Quality Assessment Based on CIEDE2000", 2012, <http://dx.doi.org/10.1155/2012/273723>.

[COMPARECODECS] Alvestrand, H., "Compare Codecs", 2015, <http://compare-codecs.appspot.com/>.

[DAALA-GIT] Xiph.Org, "Daala Git Repository", 2015, <http://git.xiph.org/?p=daala.git;a=summary>.

[I-D.ietf-netvc-requirements] Filippov, A., Norkin, A., and j. jose.roberto.alvarez@huawei.com, "Video Codec Requirements and Evaluation Methodology", draft-ietf-netvc-requirements-10 (work in progress), November 2019.

[MSSSIM] Wang, Z., Simoncelli, E., and A. Bovik, "Multi-Scale Structural Similarity for Image Quality Assessment", n.d., <http://www.cns.nyu.edu/~zwang/files/papers/msssim.pdf>.

[PSNRHVS] Egiazarian, K., Astola, J., Ponomarenko, N., Lukin, V., Battisti, F., and M. Carli, "A New Full-Reference Quality Metrics Based on HVS", 2002.

[RD_TOOL] Xiph.Org, "rd_tool", 2016, <https://github.com/tdaede/rd_tool>.

[SSIM] Wang, Z., Bovik, A., Sheikh, H., and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", 2004, <http://www.cns.nyu.edu/pub/eero/wang03-reprint.pdf>.

[TESTSEQUENCES] Daede, T., "Test Sets", n.d., <https://people.xiph.org/~tdaede/sets/>.

[VMAF] Aaron, A., Li, Z., Manohara, M., Lin, J., Wu, E., and C. Kuo, "VMAF - Video Multi-Method Assessment Fusion", 2015, <https://github.com/Netflix/vmaf>.

## Authors' Addresses

Thomas Daede
Mozilla

Email: tdaede@mozilla.com

Andrey Norkin
Netflix

Email: anorkin@netflix.com

Ilya Brailovskiy
Amazon Lab126

Email: brailovs@lab126.com

netvc
Internet-Draft
Intended status: Informational
Expires: October 26, 2017

T. Terriberry
N. Egge
Mozilla Corporation

April 24, 2017

# Coding Tools for a Next Generation Video Codec
## draft-terriberry-netvc-codingtools-02

Abstract

This document proposes a number of coding tools that could be incorporated into a next-generation video codec.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on October 26, 2017.

Table of Contents



## 1. Introduction

One of the biggest contributing factors to the success of the Internet is that the underlying protocols are implementable on a royalty-free basis. This allows them to be implemented widely and easily distributed by application developers, service operators, and end users, without asking for permission. In order to produce a next-generation video codec that is competitive with the best patent-encumbered standards, yet avoids patents which are not available on an open-source compatible, royalty-free basis, we must use old coding tools in new ways and develop new coding tools. This draft documents some of the tools we have been working on for inclusion in such a codec. This is early work, and the performance of some of these tools (especially in relation to other approaches) is not yet fully known. Nevertheless, it still serves to outline some possibilities that NETVC could consider.

## 2. Entropy Coding

The basic theory of entropy coding was well-established by the late 1970's [Pas76]. Modern video codecs have focused on Huffman codes (or "Variable-Length Codes"/VLCs) and binary arithmetic coding.

Huffman codes are limited in the amount of compression they can provide and the design flexibility they allow, but as each code word consists of an integer number of bits, their implementation complexity is very low, so they were provided at least as an option in every video codec up through H.264. Arithmetic coding, on the other hand, uses code words that can take up fractional parts of a bit, and are more complex to implement. However, the prevalence of cheap, H.264 High Profile hardware, which requires support for arithmetic coding, shows that it is no longer so expensive that a fallback VLC-based approach is required. Having a single entropy-coding method simplifies both up-front design costs and interoperability.

However, the primary limitation of arithmetic coding is that it is an inherently serial operation. A given symbol cannot be decoded until the previous symbol is decoded, because the bits (if any) that are output depend on the exact state of the decoder at the time it is decoded. This means that a hardware implementation must run at a sufficiently high clock rate to be able to decode all of the symbols in a frame. Higher clock rates lead to increased power consumption, and in some cases the entropy coding is actually becoming the limiting factor in these designs.

As fabrication processes improve, implementers are very willing to trade increased gate count for lower clock speeds. So far, most approaches to allowing parallel entropy coding have focused on splitting the encoded symbols into multiple streams that can be decoded independently. This "independence" requirement has a non-negligible impact on compression, parallelizability, or both. For example, H.264 can split frames into "slices" which might cover only a small subset of the blocks in the frame. In order to allow decoding these slices independently, they cannot use context information from blocks in other slices (harming compression). Those contexts must adapt rapidly to account for the generally small number of symbols available for learning probabilities (also harming compression). In some cases the number of contexts must be reduced to ensure enough symbols are coded in each context to usefully learn probabilities at all (once more, harming compression). Furthermore, an encoder must specially format the stream to use multiple slices per frame to allow any parallel entropy decoding at all. Encoders rarely have enough information to evaluate this "compression efficiency" vs. "parallelizability" trade-off, since they don’t generally know the limitations of the decoders for which they are encoding. That means there will be many files or streams which could have been decoded if they were encoded with different options, but which a given decoder cannot decode because of bad choices made by the encoder (at least from the perspective of that decoder). The

same set of drawbacks apply to the DCT token partitions in VP8 [RFC6386].

## 2.1. Non-binary Arithmetic Coding

Instead, we propose a very different approach: use non-binary arithmetic coding. In binary arithmetic coding, each decoded symbol has one of two possible values: 0 or 1. The original arithmetic coding algorithms allow a symbol to take on any number of possible values, and allow the size of that alphabet to change with each symbol coded. Reasonable values of N (for example, N <= 16) offer the potential for a decent throughput increase for a reasonable increase in gate count for hardware implementations.

Binary coding allows a number of computational simplifications. For example, for each coded symbol, the set of valid code points is partitioned in two, and the decoded value is determined by finding the partition in which the actual code point that was received lies. This can be determined by computing a single partition value (in both the encoder and decoder) and (in the decoder) doing a single comparison. A non-binary arithmetic coder partitions the set of valid code points into multiple pieces (one for each possible value of the coded symbol). This requires the encoder to compute two partition values, in general (for both the upper and lower bound of the symbol to encode). The decoder, on the other hand, must search the partitions for the one that contains the received code point. This requires computing at least O(log N) partition values.

However, coding a parameter with N possible values with a binary arithmetic coder requires O(log N) symbols in the worst case (the only case that matters for hardware design). Hence, this does not represent any actual savings (indeed, it represents an increase in the number of partition values computed by the encoder). In addition, there are a number of overheads that are per-symbol, rather than per-value. For example, renormalization (which enlarges the set of valid code points after partitioning has reduced it too much), carry propagation (to deal with the case where the high and low ends of a partition straddle a bit boundary), etc., are all performed on a symbol-by-symbol basis. Since a non-binary arithmetic coder codes a given set of values with fewer symbols than a binary one, it incurs these per-symbol overheads less often. This suggests that a non-binary arithmetic coder can actually be more efficient than a binary one.

## 2.2. Non-binary Context Modeling

The other aspect that binary coding simplifies is probability modeling. In arithmetic coding, the size of the sets the code points are partitioned into are (roughly) proportional to the probability of each possible symbol value. Estimating these probabilities is part of the coding process, though it can be cleanly separated from the task of actually producing the coded bits. In a binary arithmetic coder, this requires estimating the probability of only one of the two possible values (since the total probability is 1.0). This is often done with a simple table lookup that maps the old probability and the most recently decoded symbol to a new probability to use for the next symbol in the current context. The trade-off, of course, is that non-binary symbols must be "binarized" into a series of bits, and a context (with an associated probability) chosen for each one.

In a non-binary arithmetic coder, the decoder must compute at least O(log N) cumulative probabilities (one for each partition value it needs). Because these probabilities are usually not estimated directly in "cumulative" form, this can require computing (N - 1) non-cumulative probability values. Unless N is very small, these cannot be updated with a single table lookup. The normal approach is to use "frequency counts". Define the frequency of value k to be

f[k] = A*<the number of times k has been observed> + B

where A and B are parameters (usually A=2 and B=1 for a traditional Krichevsky-Trofimov estimator). The resulting probability, p[k], is given by

$$ft = \sum_{k=0}^{N-1} f[k]$$

$$p[k] = \frac{f[k]}{ft}$$

When ft grows too large, the frequencies are rescaled (e.g., halved, rounding up to prevent reduction of a probability to 0).

When ft is not a power of two, partitioning the code points requires actual divisions (see [RFC6716] Section 4.1 for one detailed example of exactly how this is done). These divisions are acceptable in an audio codec like Opus [RFC6716], which only has to code a few hundreds of these symbols per second. But video requires hundreds of

thousands of symbols per second, at a minimum, and divisions are still very expensive to implement in hardware.

There are two possible approaches to this. One is to come up with a replacement for frequency counts that produces probabilities that sum to a power of two. Some possibilities, which can be applied individually or in combination:

1. Use probabilities that are fixed for the duration of a frame. This is the approach taken by VP8, for example, even though it uses a binary arithmetic coder. In fact, it is possible to convert many of VP8's existing binary-alphabet probabilities into probabilities for non-binary alphabets, an approach that is used in the experiment presented at the end of this section.

2. Use parametric distributions. For example, DCT coefficient magnitudes usually have an approximately exponential distribution. This distribution can be characterized by a single parameter, e.g., the expected value. The expected value is trivial to update after decoding a coefficient. For example

   E[x[n+1]] = E[x[n]] + floor(C*(x[n] - E[x[n]]))

   produces an exponential moving average with a decay factor of (1 - C). For a choice of C that is a negative power of two (e.g., 1/16 or 1/32 or similar), this can be implemented with two adds and a shift. Given this expected value, the actual distribution to use can be obtained from a small set of pre-computed distributions via a lookup table. Linear interpolation between these pre-computed values can improve accuracy, at the cost of O(N) computations, but if N is kept small this is trivially parallelizable, in SIMD or otherwise.

3. Change the frequency count update mechanism so that ft is constant. This approach is described in the next section.

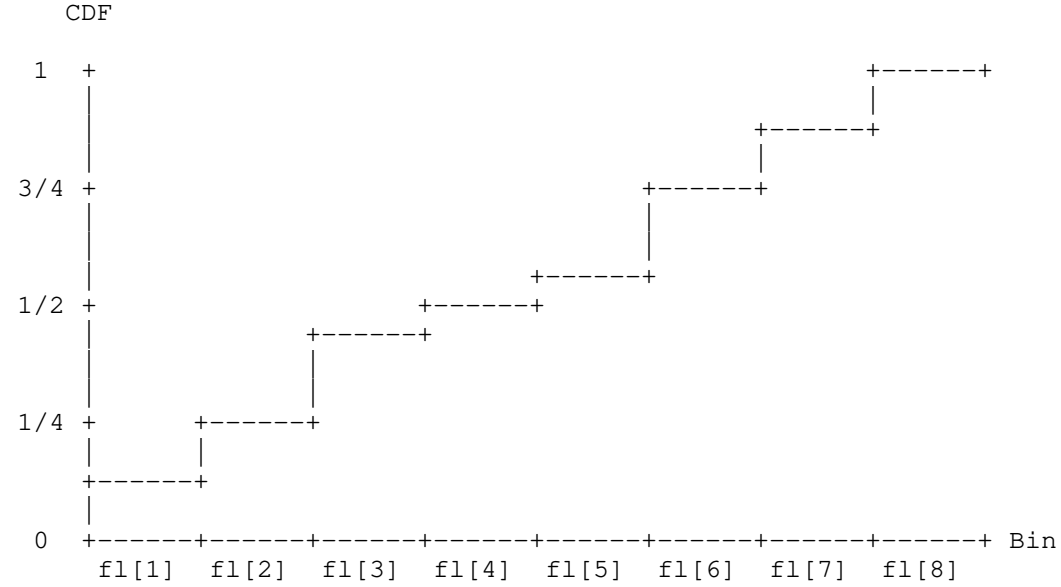
## 2.3. Dyadic Adaptation

The goal with context adaptation using dyadic probabilities is to maintain the invariant that the probabilities all sum to a power of two before and after adaptation. This can be achieved with a special update function that blends the cumulative probabilities of the current context with a cumulative distribution function where the coded symbol has probability 1.

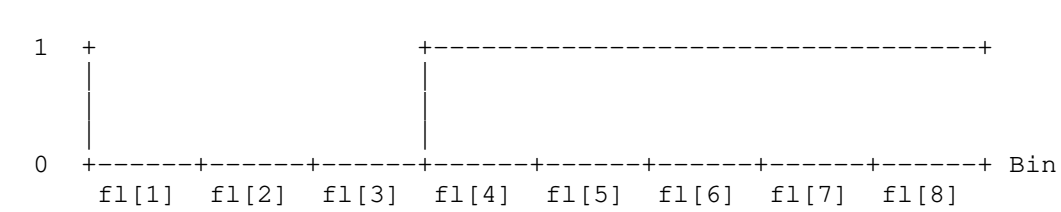
Suppose we have model for a given context that codes 8 symbols with the following probabilities:

| p[0] | p[1] | p[2] | p[3] | p[4] | p[5] | p[6] | p[7] |
|---|---|---|---|---|---|---|---|
| 1/8 | 1/8 | 3/16 | 1/16 | 1/16 | 3/16 | 1/8 | 1/8 |

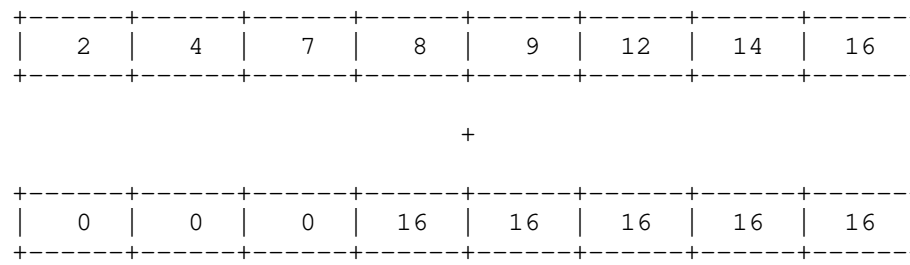
Then the cumulative distribution function is:




Suppose we code symbol 3 and wish to update the context model so that this symbol is now more likely. This can be done by blending the CDF for the current context with a CDF that has symbol 3 with likelihood 1.




Given an adaptation rate g between 0 and 1, and assuming ft = 2^4 = 16, what we are computing is:

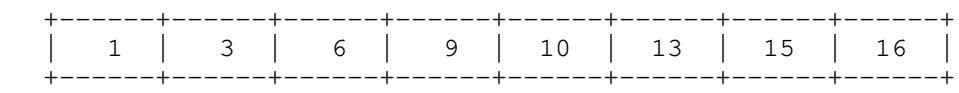



In order to prevent the probability of any one symbol from going to zero, the blending functions above and below the coded symbol are adjusted so that no adjacent cumulative probabilities are the same.

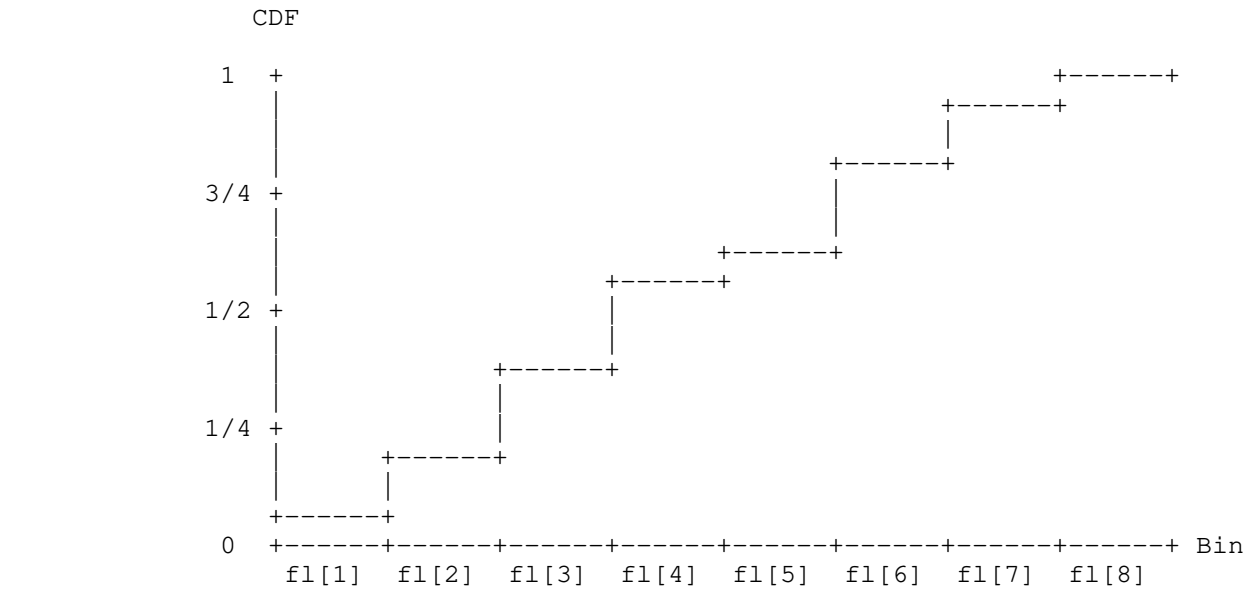
Let M be the alphabet size and 1/2^r be the adaptation rate:

```
        ( fl[i] - floor((fl[i] + 2^r - i - 1)/2^r), i <= coded symbol
fl[i] = <
        ( fl[i] - floor((fl[i] + M - i - ft)/2^r),  i > coded symbol
```

Applying these formulas to the example CDF where M = 8 with adaptation rate 1/2^16 gives the updated CDF:




Looking at the graph of the CDF we see that the likelihood for symbol 3 has gone up from 1/16 to 3/16, dropping the likelihood of all other symbols to make room.




## 2.4. Simplified Partition Function

Let the range of valid code points in the current arithmetic coder state be [L, L + R), where L is the lower bound of the range and R is the number of valid code points. The goal of the arithmetic coder is to partition this interval proportional to the probability of each symbol. When using dyadic probabilities, the partition point in the range corresponding to a given CDF value can be determined via

$$u[k] = \text{floor}\left(\frac{fl[k]*R}{ft}\right)$$

Since ft is a power of two, this may be implemented using a right shift by T bits in place of the division:

$$u[k] = (fl[k]*R) >> T$$

The latency of the multiply still dominates the hardware timing. However, we can reduce this latency by using a smaller multiply, at the cost of some accuracy in the partition. We cannot, in general, reduce the size of fl[k], since this might send a probability to zero (i.e., cause u[k] to have the same value as u[k+1]). On the other hand, we know that the top bit of R is always 1, since it gets renormalized with every symbol that is encoded. Suppose R contains 16 bits and that T is at least 8. Then we can greatly reduce the size of the multiply by using the formula

```
                ( (fl[k]*(R >> 8)) >> (T - 8), 0 <= k < M
        u[k] = <
                ( R,                           k == M
```

The special case for k == M is required because, with the general formula, u[M] no longer exactly equals R. Without the special case we would waste some amount of code space and require the decoder to check for invalid streams. This special case slightly inflates the probability of the last symbol. Unfortunately, in codecs the usual convention is that the last symbol is the least probable, while the first symbol (e.g., 0) is the most probable. That maximizes the coding overhead introduced by this approximation error. To minimize it, we instead add all of the accumulated error to the first symbol by using a variation of the above update formula:

```
                ( 0,                                                 k == 0
        u[k] = <
                ( R - (((ft - fl[k])*(R >> 8)) >> (T - 8)), 0 < k <= M
```

This also aids the software decoder search, since it can prime the search loop with the special case, instead of needing to check for it on every iteration of the loop. It is easier to incorporate into a SIMD search as well. It does, however, add two subtractions. Since the encoder always operates on the difference between two partition points, the first subtraction (involving R) can be eliminated. Similar optimizations can eliminate this subtraction in the decoder by flipping its internal state (measuring the distance of the encoder output from the top of the range instead of the bottom). To avoid the other subtraction, we can simply use "inverse CDFs" that natively store ifl[k] = (ft - fl[k]) instead of fl[k]. This produces the following partition function:

```
                    ( R,                               k == 0
        R - u[k] = <
                    ( (ifl[k]*(R >> 8)) >> (T - 8), 0 < k <= M
```

The reduction in hardware latency can be as much as 20%, and the impact on area is even larger. The overall software complexity overhead is minimal, and the coding efficiency overhead due to the approximation is about 0.02%. We could have achieved the same efficiency by leaving the special case on the last symbol and reversing the alphabet instead of inverting the probabilities. However, reversing the alphabet at runtime would have required an extra subtraction (or more general re-ordering requires a table lookup). That may be avoidable in some cases, but only by propagating the reordering alphabet outside of the entropy coding machinery, requiring changes to every coding tool and potentially leading to confusion. CDFs, on the other hand, are already a

somewhat abstract representation of the underlying probabilities used for computational efficiency reasons. Generalizing these to "inverse CDFs" is a straightforward change that only affects probability initialization and adaptation, without impacting the design of other coding tools.

### 2.5. Context Adaptation

The dyadic adaptation scheme described in Section 2.3 implements a low-complexity IIR filter for the steady-state case where we only want to adapt the context CDF as fast as the 1/2^r adaptation rate. In many cases, for example when coding symbols at the start of a video frame, only a limited number of symbols have been seen per context. Using this steady-state adaptation scheme risks adapting too slowly and spending too many bits to code symbols with incorrect probability estimates. In other video codecs, this problem is reduced by either implicitly or explicitly allowing for mechanisms to set the initial probability models for a given context.

#### 2.5.1. Implicit Adaptation

One implicit way to use default probabilities is to simply require as a normative part of the decoder that some specific CDFs are used to initialize each context. A representative set of inputs is run through the encoder and a frequency based probability model is computed and reloaded at the start of every frame. This has the advantage of having zero bitstream overhead and is optimal for certain stationary symbols. However for other non-stationary symbols, or highly content dependent contexts where the sample input is not representative, this can be worse than starting with a flat distribution as it now takes even longer to adapt to the steady-state. Moreover the amount of hardware area required to store initial probability tables for each context goes up with the number of contexts in the codec.

Another implicit way to deal with poor initial probabilities is through backward adaptation based on the probability estimates from the previous frame. After decoding a frame, the adapted CDFs for each context are simply kept as-is and not reset to their defaults. This has the advantage of having no bitstream overhead, and tracking to certain content types closely as we expect frames with similar content at similar rates, to have well correlated CDFs. However, this only works when we know there will be no bitstream errors due to the transport layer, e.g., TCP or HTTP. In low delay use cases (video on demand, live streaming, video conferencing), implicit backwards adaptation is avoided as it risks desynchronizing the entropy decoder state and permanently losing the video stream.

### 2.5.2. Explicit Adaptation

For codecs that include the ability to update the probability models in the bitstream, it is possible to explicitly signal a starting CDF. The previously described implicit backwards adaptation is now possible by simply explicitly coding a probability update for each frame. However, the cost of signaling the updated CDF must be overcome by the savings from coding with the updated CDF. Blindly updating all contexts per frame may work at high rates where the size of the CDFs is small relative to the coded symbol data. However at low rates, the benefit of using more accurate CDFs is quickly overcome by the cost of coding them, which increases with the number of contexts.

More sophisticated encoders can compute the cost of coding a probability update for a given context, and compare it to the size reduction achieved by coding symbols with this context. Here all symbols for a given frame (or tile) are buffered and not serialized by the entropy coder until the end of the frame (or tile) is reached. Once the end of the entropy segment has been reached, the cost in bits for coding symbols with both the default probabilities and the proposed updated probabilities can be measured and compared. However, note that with the symbols already buffered, rather than consider the context probabilities from the previous frame, a simple frequency based probability model can be computed and measured. Because this probability model is computed based on the symbols we are about to code this technique is called forward adaptation. If the cost in bits to signal and code with this new probability model is less than that of using the default then it is used. This has the advantage of only ever coding a probability update if it is an improvement and producing a bitstream that is robust to errors, but requires an entire entropy segments worth of symbols be cached.

### 2.5.3. Early Adaptation

We would like to take advantage of the low-cost multi-symbol CDF adaptation described in Section 2.3 without in the broadest set of use cases. This means the initial probability adaptation scheme should support low-delay, error-resilient streams that efficiently implemented in both hardware and software. We propose an early adaptation scheme that supports this goal.

At the beginning of a frame (or tile), all CDFs are initialized to a flat distribution. For a given multi-symbol context with M potential symbols, assume that the initial dyadic CDF is initialized so that each symbol has probability 1/M. For the first M coded symbols, the CDF is updated as follows:

```
a[c,M] = ft/(M + c)

        ( fl[i] - floor((fl[i] - i)*a/ft),          i <= coded symbol
fl[i] = <
        ( fl[i] - floor((fl[i] + M - i - ft)*a/ft), i > coded symbol
```

where c goes from 0 to M-1 and is the running count of the number of symbols coded with this CDF. Note that for a fixed CDF precision (ft is always a power of two) and a maximum number of possible symbols M, the values of a[c,M] can be stored in a M*(M+1)/2 element table, which is 136 entries when M = 16.

## 2.6. Simple Experiment

As a simple experiment to validate the non-binary approach, we compared a non-binary arithmetic coder to the VP8 (binary) entropy coder. This was done by instrumenting vp8_treed_read() in libvpx to dump out the symbol decoded and the associated probabilities used to decode it. This data only includes macroblock mode and motion vector information, as the DCT token data is decoded with custom inline functions, and not vp8_treed_read(). This data is available at [1]. It includes 1,019,670 values encode using 2,125,995 binary symbols (or 2.08 symbols per value). We expect that with a conscious effort to group symbols during the codec design, this average could easily be increased.

We then implemented both the regular VP8 entropy decoder (in plain C, using all of the optimizations available in libvpx at the time) and a multisymbol entropy decoder (also in plain C, using similar optimizations), which encodes each value with a single symbol. For the decoder partition search in the non-binary decoder, we used a simple for loop (O(N) worst-case), even though this could be made constant-time and branchless with a few SIMD instructions such as (on x86) PCMPGTW, PACKUSWB, and PMOVMASKB followed by BSR. The source code for both implementations is available at [2] (compile with -DEC_BINARY for the binary version and -DEC_MULTISYM for the non-binary version).

The test simply loads the tokens, and then loops 1024 times encoding them using the probabilities provided, and then decoding them. The loop was added to reduce the impact of the overhead of loading the data, which is implemented very inefficiently. The total runtime on a Core i7 from 2010 is 53.735 seconds for the binary version, and 27.937 seconds for the non-binary version, or a 1.92x improvement. This is very nearly equal to the number of symbols per value in the binary coder, suggesting that the per-symbol overheads account for the vast majority of the computation time in this implementation.

## 3. Reversible Integer Transforms

Integer transforms in image and video coding date back to at least 1969 [PKA69]. Although standards such as MPEG2 and MPEG4 Part 2 allow some flexibility in the transform implementation, implementations were subject to drift and error accumulation, and encoders had to impose special macroblock refresh requirements to avoid these problems, not always successfully. As transforms in modern codecs only account for on the order of 10% of the total decoder complexity, and, with the use of weighted prediction with gains greater than unity and intra prediction, are far more susceptible to drift and error accumulation, it no longer makes sense to allow a non-exact transform specification.

However, it is also possible to make such transforms "reversible", in the sense that applying the inverse transform to the result of the forward transform gives back the original input values, exactly. This gives a lossy codec, which normally quantizes the coefficients before feeding them into the inverse transform, the ability to scale all the way to lossless compression without requiring any new coding tools. This approach has been used successfully by JPEG XR, for example [TSSRM08].

Such reversible transforms can be constructed using "lifting steps", a series of shear operations that can represent any set of plane rotations, and thus any orthogonal transform. This approach dates back to at least 1992 [BE92], which used it to implement a four-point 1-D Discrete Cosine Transform (DCT). Their implementation requires 6 multiplications, 10 additions, 2 shifts, and 2 negations, and produces output that is a factor of sqrt(2) larger than the orthonormal version of the transform. The expansion of the dynamic range directly translates into more bits to code for lossless compression. Because the least significant bits are usually very nearly random noise, this scaling increases the coding cost by approximately half a bit per sample.

### 3.1. Lifting Steps

To demonstrate the idea of lifting steps, consider the two-point transform

$$\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$$

This can be implemented up to scale via

$$y0 = x0 + x1$$

$$y1 = 2*x1 - y0$$

and reversed via

$$x1 = (y0 + y1) >> 1$$

$$x0 = y0 - x1$$

Both y0 and y1 are too large by a factor of sqrt(2), however.

It is also possible to implement any rotation by an angle t, including the orthonormal scale factor, by decomposing it into three steps:

$$u0 = x0 + \frac{\cos(t) - 1}{\sin(t)} * x1$$

$$y1 = x1 + \sin(t)*u0$$

$$y0 = u0 + \frac{\cos(t) - 1}{\sin(t)} * y1$$

By letting t=-pi/4, we get an implementation of the first transform that includes the scaling factor. To get an integer approximation of this transform, we need only replace the transcendental constants by fixed-point approximations:

$$u0 = x0 + ((27*x1 + 32) >> 6)$$

$$y1 = x1 - ((45*u0 + 32) >> 6)$$

$$y0 = u0 + ((27*y1 + 32) >> 6)$$

This approximation is still perfectly reversible:

$$u0 = y0 - ((27*y1 + 32) >> 6)$$

$$x1 = y1 + ((45*u0 + 32) >> 6)$$

$$x0 = u0 - ((27*x1 + 32) >> 6)$$

Each of the three steps can be implemented using just two ARM instructions, with constants that have up to 14 bits of precision (though using fewer bits allows more efficient hardware

implementations, at a small cost in coding gain). However, it is still much more complex than the first approach.

We can get a compromise with a slight modification:

```
y0 = x0 + x1

y1 = x1 - (y0 >> 1)
```

This still only implements the original orthonormal transform up to scale. The y0 coefficient is too large by a factor of sqrt(2) as before, but y1 is now too small by a factor of sqrt(2). If our goal is simply to (optionally quantize) and code the result, this is good enough. The different scale factors can be incorporated into the quantization matrix in the lossy case, and the total expansion is roughly equivalent to that of the orthonormal transform in the lossless case. Plus, we can perform each step with just one ARM instruction.

However, if instead we want to apply additional transformations to the data, or use the result to predict other data, it becomes much more convenient to have uniformly scaled outputs. For a two-point transform, there is little we can do to improve on the three-multiplications approach above. However, for a four-point transform, we can use the last approach and arrange multiple transform stages such that the "too large" and "too small" scaling factors cancel out, producing a result that has the true, uniform, orthonormal scaling. To do this, we need one more tool, which implements the following transform:

```
                 ___
[ y0 ]     / 1   [ cos(t) -sin(t) ] [ 1  0 ] [ x0 ]
[    ] =  / --- [                ] [     ] [    ]
[ y1 ]   v   2   [ sin(t)  cos(t) ] [ 0  2 ] [ x1 ]
```

This takes unevenly scaled inputs, rescales them, and then rotates them. Like an ordinary rotation, it can be reduced to three lifting steps:

$$u0 = x0 + \frac{2*\cos(t) - \overline{v2}}{\sin(t)} * x1$$

$$y1 = x1 + \sqrt{\frac{1}{v}} * \frac{1}{2} * \sin(t)*u0$$

$$y0 = u0 + \frac{\cos(t) - \overline{v2}}{\sin(t)} * y1$$

As before, the transcendental constants may be replaced by fixed-point approximations without harming the reversibility property.

### 3.2. 4-Point Transform

Using the tools from the previous section, we can design a reversible integer four-point DCT approximation with uniform, orthonormal scaling. This requires 3 multiplies, 9 additions, and 2 shifts (not counting the shift and rounding offset used in the fixed-point multiplies, as these are built into the multiplier). This is significantly cheaper than the [BE92] approach, and the output scaling is smaller by a factor of sqrt(2), saving half a bit per sample in the lossless case. By comparison, the four-point forward DCT approximation used in VP9, which is not reversible, uses 6 multiplies, 6 additions, and 2 shifts (counting shifts and rounding offsets which cannot be merged into a single multiply instruction on ARM). Four of its multipliers also require 28-bit accumulators, whereas this proposal can use much smaller multipliers without giving up the reversibility property. The total dynamic range expansion is 1 bit: inputs in the range [-256,255) produce transformed values in the range [-512,510). This is the smallest dynamic range expansion possible for any reversible transform constructed from mostly-linear operations. It is possible to make reversible orthogonal transforms with no dynamic range expansion by using "piecewise-linear" rotations [SLD04], but each step requires a large number of operations in a software implementation.

Pseudo-code for the forward transform follows:

```
Input:  x0, x1, x2, x3
Output: y0, y1, y2, y3
/* Rotate (x3, x0) by -pi/4, asymmetrically scaled output. */
t3  = x0 - x3
t0  = x0 - (t3 >> 1)
/* Rotate (x1, x2) by pi/4, asymmetrically scaled output. */
t2  = x1 + x2
t2h = t2 >> 1
t1  = t2h - x2
/* Rotate (t2, t0) by -pi/4, asymmetrically scaled input. */
y0  = t0 + t2h
y2  = y0 - t2
/* Rotate (t3, t1) by 3*pi/8, asymmetrically scaled input. */
t3  = t3 - (45*t1 + 32 >> 6)
y1  = t1 + (21*t3 + 16 >> 5)
y3  = t3 - (71*y1 + 32 >> 6)
```

Even though there are three asymmetrically scaled rotations by pi/4, by careful arrangement we can share one of the shift operations (to help software implementations: shifts by a constant are basically free in hardware). This technique can be used to even greater effect in larger transforms.

The inverse transform is constructed by simply undoing each step in turn:

```
Input:  y0, y1, y2, y3
Output: x0, x1, x2, x3
/* Rotate (y3, y1) by -3*pi/8, asymmetrically scaled output. */
t3  = y3 + (71*y1 + 32 >> 6)
t1  = y1 - (21*t3 + 16 >> 5)
t3  = t3 + (45*t1 + 32 >> 6)
/* Rotate (y2, y0) by pi/4, asymmetrically scaled output. */
t2  = y0 - y2
t2h = t2 >> 1
t0  = y0 - t2h
/* Rotate (t1, t2) by -pi/4, asymmetrically scaled input. */
x2  = t2h - t1
x1  = t2 - x2
/* Rotate (x3, x0) by pi/4, asymmetrically scaled input. */
x0  = t0 - (t3 >> 1)
x3  = x0 - t3
```

Although the right shifts make this transform non-linear, we can compute "basis functions" for it by sending a vector through it with a single value set to a large constant (256 was used here), and the rest of the values set to zero. The true basis functions for a four-point DCT (up to five digits) are

```
[ y0 ]   [ 0.50000  0.50000  0.50000  0.50000 ] [ x0 ]
[ y1 ] = [ 0.65625  0.26953 -0.26953 -0.65625 ] [ x1 ]
[ y2 ]   [ 0.50000 -0.50000 -0.50000  0.50000 ] [ x2 ]
[ y3 ]   [ 0.27344 -0.65234  0.65234 -0.27344 ] [ x3 ]
```

The corresponding basis functions for our reversible, integer DCT, computed using the approximation described above, are

```
[ y0 ]   [ 0.50000  0.50000  0.50000  0.50000 ] [ x0 ]
[ y1 ] = [ 0.65328  0.27060 -0.27060 -0.65328 ] [ x1 ]
[ y2 ]   [ 0.50000 -0.50000 -0.50000  0.50000 ] [ x2 ]
[ y3 ]   [ 0.27060 -0.65328  0.65328 -0.27060 ] [ x3 ]
```

The mean squared error (MSE) of the output, compared to a true DCT, can be computed with some assumptions about the input signal. Let G be the true DCT basis and G' be the basis for our integer approximation (computed as described above). Then the error in the transformed results is

```
e = G.x - G'.x = (G - G').x = D.x
```

where D = (G - G') . The MSE is then [Que98]

$$\frac{1}{N} * E[e^T.e] = \frac{1}{N} * E[x^T.D^T.D.x]$$

$$= \frac{1}{N} * E[tr(D.x.x^T.D^T)]$$

$$= \frac{1}{N} * E[tr(D.Rxx.D^T)]$$

where Rxx is the autocorrelation matrix of the input signal. Assuming the input is a zero-mean, first-order autoregressive (AR(1)) process gives an autocorrelation matrix of

$$Rxx[i,j] = rho^{|i - j|}$$

for some correlation coefficient rho. A value of rho = 0.95 is typical for image compression applications. Smaller values are more normal for motion-compensated frame differences, but this makes surprisingly little difference in transform design. Using the above procedure, the theoretical MSE of this approximation is 1.230E-6, which is below the level of the truncation error introduced by the

right shift operations. This suggests the dynamic range of the input would have to be more than 20 bits before it became worthwhile to increase the precision of the constants used in the multiplications to improve accuracy, though it may be worth using more precision to reduce bias.

### 3.3. Larger Transforms

The same techniques can be applied to construct a reversible eight-point DCT approximation with uniform, orthonormal scaling using 15 multiplies, 31 additions, and 5 shifts. It is possible to reduce this to 11 multiplies and 29 additions, which is the minimum number of multiplies possible for an eight-point DCT with uniform scaling [LLM89], by introducing a scaling factor of sqrt(2), but this harms lossless performance. The dynamic range expansion is 1.5 bits (again the smallest possible), and the MSE is 1.592E-06. By comparison, the eight-point transform in VP9 uses 12 multiplications, 32 additions, and 6 shifts.

Similarly, we have constructed a reversible sixteen-point DCT approximation with uniform, orthonormal scaling using 33 multiplies, 83 additions, and 16 shifts. This is just 2 multiplies and 2 additions more than the (non-reversible, non-integer, but uniformly scaled) factorization in [LLM89]. By comparison, the sixteen-point transform in VP9 uses 44 multiplies, 88 additions, and 18 shifts. The dynamic range expansion is only 2 bits (again the smallest possible), and the MSE is 1.495E-5.

We also have a reversible 32-point DCT approximation with uniform, orthonormal scaling using 87 multiplies, 215 additions, and 38 shifts. By comparison, the 32-point transform in VP9 uses 116 multiplies, 194 additions, and 66 shifts. Our dynamic range expansion is still the minimal 2.5 bits, and the MSE is 8.006E-05

Code for all of these transforms is available in the development repository listed in Section 4.

### 3.4. Walsh-Hadamard Transforms

These techniques can also be applied to constructing Walsh-Hadamard Transforms, another useful transform family that is cheaper to implement than the DCT (since it requires no multiplications at all). The WHT has many applications as a cheap way to approximately change the time and frequency resolution of a set of data (either individual bands, as in the Opus audio codec, or whole blocks). VP9 uses it as a reversible transform with uniform, orthonormal scaling for lossless coding in place of its DCT, which does not have these properties.

Applying a 2x2 WHT to a block of 2x2 inputs involves running a 2-point WHT on the rows, and then another 2-point WHT on the columns. The basis functions for the 2-point WHT are, up to scaling, [1, 1] and [1, -1]. The four variations of a two-step lifer given in Section 3.1 are exactly the lifting steps needed to implement a 2x2 WHT: two stages that produce asymmetrically scaled outputs followed by two stages that consume asymmetrically scaled inputs.

```
Input:  x00, x01, x10, x11
Output: y00, y01, y10, y11
/* Transform rows */
t1 = x00 - x01
t0 = x00 - (t1 >> 1) /* == (x00 + x01)/2 */
t2 = x10 + x11
t3 = (t2 >> 1) - x11 /* == (x10 - x11)/2 */
/* Transform columns */
y00 = t0 + (t2 >> 1) /* == (x00 + x01 + x10 + x11)/2 */
y10 = y00 - t2       /* == (x00 + x01 - x10 - x11)/2 */
y11 = (t1 >> 1) - t3 /* == (x00 - x01 - x10 + x11)/2 */
y01 = t1 - y11       /* == (x00 - x01 + x10 - x11)/2 */
```

By simply re-ordering the operations, we can see that there are two shifts that may be shared between the two stages:

```
Input:  x00, x01, x10, x11
Output: y00, y01, y10, y11
t1 = x00 - x01
t2 = x10 + x11
t0 = x00 - (t1 >> 1) /* == (x00 + x01)/2 */
y00 = t0 + (t2 >> 1) /* == (x00 + x01 + x10 + x11)/2 */
t3 = (t2 >> 1) - x11 /* == (x10 - x11)/2 */
y11 = (t1 >> 1) - t3 /* == (x00 - x01 - x10 + x11)/2 */
y10 = y00 - t2       /* == (x00 + x01 - x10 - x11)/2 */
y01 = t1 - y11       /* == (x00 - x01 + x10 - x11)/2 */
```

By eliminating the double-negation of x11 and re-ordering the additions to it, we can see even more operations in common:

```
Input:  x00, x01, x10, x11
Output: y00, y01, y10, y11
t1 = x00 - x01
t2 = x10 + x11
t0 = x00 - (t1 >> 1) /* == (x00 + x01)/2 */
y00 = t0 + (t2 >> 1) /* == (x00 + x01 + x10 + x11)/2 */
t3 = x11 + (t1 >> 1) /* == x11 + (x00 - x01)/2 */
y11 = t3 - (t2 >> 1) /* == (x00 - x01 - x10 + x11)/2 */
y10 = y00 - t2       /* == (x00 + x01 - x10 - x11)/2 */
y01 = t1 - y11       /* == (x00 - x01 + x10 - x11)/2 */
```

Simplifying further, the whole transform may be computed with just 7 additions and 1 shift:

```
Input:  x00, x01, x10, x11
Output: y00, y01, y10, y11
t1 = x00 - x01
t2 = x10 + x11
t4 = (t2 - t1) >> 1 /* == (-x00 + x01 + x10 + x11)/2 */
y00 = x00 + t4      /* ==  (x00 + x01 + x10 + x11)/2 */
y11 = x11 - t4      /* ==  (x00 - x01 - x10 + x11)/2 */
y10 = y00 - t2      /* ==  (x00 + x01 - x10 - x11)/2 */
y01 = t1 - y11      /* ==  (x00 - x01 + x10 - x11)/2 */
```

This is a significant savings over other approaches described in the literature, which require 8 additions, 2 shifts, and 1 negation [FOIK99] (37.5% more operations), or 10 additions, 1 shift, and 2 negations [TSSRM08] (62.5% more operations). The same operations can be applied to compute a 4-point WHT in one dimension. This implementation is used in this way in VP9's lossless mode. Since larger WHTs may be trivially factored into multiple smaller WHTs, the same approach can implement a reversible, orthonormally scaled WHT of any size (2**N)x(2**M), so long as (N + M) is even.

## 4. Development Repository

The tools presented here were developed as part of Xiph.Org's Daala project. They are available, along with many others in greater and lesser states of maturity, in the Daala git repository at [3]. See [4] for more information.

## 5. IANA Considerations

This document has no actions for IANA.

## 6. Acknowledgments

Thanks to Nathan Egge, Gregory Maxwell, and Jean-Marc Valin for their assistance in the implementation and experimentation, and in preparing this draft.

## 7. References

### 7.1. Informative References

[RFC6386] Bankoski, J., Koleszar, J., Quillio, L., Salonen, J., Wilkins, P., and Y. Xu, "VP8 Data Format and Decoding Guide", RFC 6386, November 2011.

[RFC6716] Valin, JM., Vos, K., and T. Terriberry, "Definition of the Opus Audio Codec", RFC 6716, September 2012.

[BE92] Bruekers, F. and A. van den Enden, "New Networks for Perfect Inversion and Perfect Reconstruction", IEEE Journal on Selected Areas in Communication 10(1):129--137, January 1992.

[FOIK99] Fukuma, S., Oyama, K., Iwahashi, M., and N. Kambayashi, "Lossless 8-point Fast Discrete Cosine Transform Using Lossless Hadamard Transform", Technical Report The Institute of Electronics, Information, and Communication Engineers of Japan, October 1999.

[LLM89] Loeffler, C., Ligtenberg, A., and G. Moschytz, "Practical Fast 1-D DCT Algorithms with 11 Multiplications", Proc. Acoustics, Speech, and Signal Processing (ICASSP'89) vol. 2, pp. 988--991, May 1989.

[Pas76] Pasco, R., "Source Coding Algorithms for Fast Data Compression", Ph.D. Thesis Dept. of Electrical Engineering, Stanford University, May 1976.

[PKA69] Pratt, W., Kane, J., and H. Andrews, "Hadamard Transform Image Coding", Proc. IEEE 57(1):58--68, Jan 1969.

[Que98] de Queiroz, R., "On Unitary Transform Approximations", IEEE Signal Processing Letters 5(2):46--47, Feb 1998.

[SLD04] Senecal, J., Lindstrom, P., and M. Duchaineau, "An Improved N-Bit to N-Bit Reversible Haar-Like Transform", Proc. of the 12th Pacific Conference on Computer Graphics and Applications (PG'04) pp. 371--380, October 2004.

[TSSRM08] Tu, C., Srinivasan, S., Sullivan, G., Regunathan, S., and H. Malvar, "Low-complexity Hierarchical Lapped Transform for Lossy-to-Lossless Image Coding in JPEG XR/HD Photo", Applications of Digital Image Processing XXXI vol 7073, August 2008.

## 7.2. URIs

[1] https://people.xiph.org/~tterribe/daala/ec_test0/ec_tokens.txt

[2] https://people.xiph.org/~tterribe/daala/ec_test0/ec_test.c

[3] https://git.xiph.org/daala.git

[4] https://xiph.org/daala/

Authors' Addresses

Timothy B. Terriberry
Mozilla Corporation
331 E. Evelyn Avenue
Mountain View, CA 94041
USA

Phone: +1 650 903-0800
Email: tterribe@xiph.org

Nathan E. Egge
Mozilla Corporation
331 E. Evelyn Avenue
Mountain View, CA 94041
USA

Phone: +1 650 903-0800
Email: negge@xiph.org