Network Working Group
Internet-Draft
Intended status: Informational
Expires: September 01, 2016

T. Daede
Mozilla
A. Norkin
Netflix

February 29, 2016

Video Codec Testing and Quality Measurement
draft-ietf-netvc-testing-01

Abstract

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

Status of This Memo

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 and good decisions.

2. Subjective quality tests

Subjective testing is the preferable method of testing video codecs.

Because the IETF does not have testing resources of its own, it has to rely on the resources of its participants. For this reason, even

if the group agrees that a particular test is important, if no one volunteers to do it, or if volunteers do not complete it in a timely fashion, then that test should be discarded. This ensures that only important tests be done in particular, the tests that are important to participants.

2.1. Still Image Pair Comparison

A simple way to determine superiority of one compressed image over another is to visually compare two compressed images, and have the viewer judge which one has a higher quality. This is mainly used for rapid comparisons during development. 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.

2.2. Subjective viewing test

A subjective 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 could be (depending on the test procedure) mean opinion scores (MOS) or differential mean opinion scores (DMOS). Normally, confidence intervals are also calculated to judge whether the difference between two encodings is statistically significant.

2.3. Expert viewing

An expert viewing test can be performed in the case when an answer to a particular question should be found. An example of such test can be a test involving video coding experts on evaluation of a particular problem, for example such as comparing the results of two de-ringing filters. Depending on what information is sought, the appropriate test procedure can be chosen.

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].

All of the metrics described in this document are to be applied to the luma plane only. In addition, they are single frame metrics. When applied to the video, the scores of each frame are averaged to create the final score.

Codecs are allowed to internally use downsampling, but must include a normative upsampler, so that the metrics run at the same resolution as the source video. In addition, some metrics, such as PSNR and FASTSSIM, have poor behavior on downsampled images, so it must be noted in test results if downsampling is in effect.

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 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. [PSNRHVS] 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.

For the metric to appear more linear on BD-rate curves, 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].

3.6. Fast Multi-Scale SSIM

Fast MS-SSIM is a modified implementation of MS-SSIM which operates on a limited number of scales and with modified weights [FASTSSIM]. The final score is converted to decibels in the same manner as SSIM.

3.7. 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.

3.8. 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 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.

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. Bjontegaard

The Bjontegaard rate difference, also known as BD-rate, allows the comparison of two different codecs based on a metric. This is commonly done by fitting a curve to each set of data points on the plot of bitrate versus metric score, and then computing the difference in area between each of the curves. A cubic polynomial fit is common, but will be overconstrained with more than four samples. For higher accuracy, at least 10 samples and a cubic spline fit should be used. In addition, if using a truncated BD-rate curve, there should be at least 4 samples within the point of interest.

4.3. Ranges

The curve is split into three regions, for low, medium, and high bitrate. The ranges are defined as follows:

- Low bitrate: 0.005 - 0.02 bpp

- Medium bitrate: 0.02 - 0.06 bpp

- High bitrate: 0.06 - 0.2 bpp

Bitrate can be calculated from bits per pixel (bpp) as follows:

bitrate = bpp * width * height * framerate

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].

- Still images are useful when comparing intra coding performance. Xiph.org has four sets of lossless, one megapixel images that have been converted into YUV 4:2:0 format. There are four sets that can be used:

  * subset1 (50 images)

  * subset2 (50 images)

  * subset3 (1000 images)

  * subset4 (1000 images)

- video-hd-3, a set that consists of 1920x1080 clips from [DERFVIDEO] (1500 frames total)

- vc-360p-1, a low quality video conferencing set (2700 frames total)

- vc-720p-1, a high quality video conferencing set (2750 frames total)

- netflix-4k-1, a cinematic 4K video test set (2280 frames total)

- netflix-2k-1, a 2K scaled version of netflix-4k-1 (2280 frames total)

- twitch-1, a game sequence set (2280 frames total)

5.3. Operating Points

Two 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.

5.3.1. Common settings

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

- vp10: -codec=vp10 -ivf -frame-parallel=0 -tile-columns=0 -cpu-used=0 -threads=1

5.3.2. 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

- vp10: -end-usage=q -cq-level=x -lag-in-frames=25 -auto-alt-ref=2

5.3.3. 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

- vp10: -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 the AreWeCompressedYet

tool [AWCY]. Because of computational constraints, several levels of testing are specified.

6.1. Regression tests

Regression tests run on a small number of short sequences. 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.

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. If the option for the objective performance test is chosen, wide range and full length simulations are run on the site and the results (including all the objective metrics) are generated.

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. Informative References

[AWCY] Xiph.Org, "Are We Compressed Yet?", 2015, <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>.

[DERFVIDEO]

Terriberry, T., "Xiph.org Video Test Media", n.d., <https://media.xiph.org/video/derf/>.

[FASTSSIM] Chen, M. and A. Bovik, "Fast structural similarity index algorithm", 2010, <http://live.ece.utexas.edu/publications/2011/chen_rtip_2011.pdf>.

[L1100] Bossen, F., "Common test conditions and software reference configurations", JCTVC L1100, 2013, <http://phenix.int-evry.fr/jct/>.

[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.

[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>.

[STEAM] Valve Corporation, "Steam Hardware & Software Survey: June 2015", June 2015, <http://store.steampowered.com/hwsurvey>.

[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, "Challenges in cloud based ingest and encoding for high quality streaming media", 2015, <http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7351097>.

Authors' Addresses

Thomas Daede
Mozilla

Email: tdaede@mozilla.com

Andrey Norkin
Netflix

Email: anorkin@netflix.com