Network Working Group
Internet-Draft
Intended status: Informational
Expires: June 02, 2016

T. Daede
Mozilla

November 30, 2015

Video Codec Testing and Quality Measurement
draft-ietf-netvc-testing-00

Abstract

This document describes guidelines and procedures for evaluating an internet video codec specified at the IETF. 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 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 June 02, 2016.

Table of Contents

1. Introduction

When developing an internet 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 Metrics

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.

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. 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)

which is the method used in the dump_psnr.c reference implementation.

3.2. 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.

3.3. 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)

3.4. Fast Multi-Scale SSIM

Multi-Scale SSIM is SSIM extended to multiple window sizes [MSSSIM]. This is implemented in the Fast implementation by downscaling the image a number of times, and computing SSIM over the same number of pixels, then averaging the SSIM scores together [FASTSSIM]. The final score is converted to decibels in the same manner as SSIM.

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 clarity, the X axis bitrate is always graphed in the log domain. The Y axis metric should also be chosen so that the graph is approximately linear. For metrics such as PSNR and PSNR-HVS, the metric result is already in the log domain and is left as-is. SSIM and FASTSSIM, on the other hand, return a result between 0 and 1. To create more linear graphs, this result is converted to a value in decibels:

-1 * log10 (1 - SSIM)

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 linear piecewise 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.

The JCT-VC standards organization includes a set of standard test clips for video codec testing, and parameters to run the clips with [L1100]. These clips are not publicly available, but are very useful for comparing to published results.

Xiph publishes a variety of test clips collected from various sources.

The Blender Open Movie projects provide a large test base of lossless cinematic test material. The lossless sources are available, hosted on Xiph.

5.2. Test Sets

Sources are divided into several categories to test different scenarios the codec will be required to operate in. For easier comaprison, 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.

- 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-2, a set that consists of the following 1920x1080 clips from [DERFVIDEO], cropped to 50 frames (and converted to 4:2:0 if necessary)
  - aspen
  - blue_sky
  - crowd_run
  - ducks_take_off
  - factory
  - life
  - old_town_cross
  - park_joy
  - pedestrian_area
  - red_kayak
  - riverbed

    * rush_hour

    * station2

- A video conferencing test set, with 1280x720 content at 60 frames per second. Unlike other sets, the videos in this set are 10 seconds long.

    * TBD

- Game streaming content: 1920x1080, 60 frames per second, 4:2:0 chroma subsampling. 1080p is chosen as it is currently the most common gaming monitor resolution [STEAM]. All clips should be two seconds long.

    * TBD

- Screensharing content is low framerate, high resolution content typical of a computer desktop.

    * screenshots - desktop screenshots of various resolutions, with 4:2:0 subsampling

    * Video sets TBD

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. 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

- libvpx: -codec=vp9 -end-usage=q -cq-level=x -lag-in-frames=25 -auto-alt-ref=1

5.3.2. 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
- libvpx: -codec=vp9 -end-usage=q -cq-level=x -lag-in-frames=0 -auto-alt-ref=0

5.3.3. Constrained Low Latency

The encoder is given one parameter, which is absolute bitrate. No frame delay, buffering, or lookahead is allowed. The maximum achieved bitrate deviation from the supplied parameter is determined by a buffer model:

- The buffer starts out empty.
- After each frame is encoded, the buffer is filled by the number of bits spent for the frame.
- The buffer is then emptied by (bitrate * frame duration) bits.
- The buffer fill level is checked. If it is over the limit, the test is considered a failure.

The buffer size limit is defined by the bitrate target * 0.3 seconds.

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 computer for fast results [AWCY].

7. Informative References

[AWCY] Xiph.Org, "Are We Compressed Yet?", 2015, <https://arewecompressedyet.com/>.

[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>.

Author's Address

Thomas Daede
Mozilla

Email: tdaede@mozilla.com