VQEG's Implementer's Guide to Video Quality Metrics (IGVQM) project

By: Ioannis Katsavounidis, Research Scientist, Video Infrastructure, Facebook
Interim VQEG-IGVQM chair

With additional input from:

- Chulhee Lee (chulhee@yonsei.ac.kr)
- Werner Robitza (werner.robitza@gmail.com)
- Tony Orchard: anthony.k.orchard@intel.com
- Lucjan Janowski: janowski@kt.agh.edu.pl
- Zachary Plovanic: zachary.plovanic@intel.com
- Vittorio Baroncini: baroncini@gmx.com
- Nabajeet Barman (n.barman@ieee.org)
- Mikołaj Leszczuk (mikolaj.leszczuk@agh.edu.pl)
- Zhenzhong Chen (zzchen@whu.edu.cn)
- Silvio Borer: silvio.borer@rohde-schwarz.com
- Jesús Gutiérrez: jesus.gutierrez@upm.es
- Lukas Krasula (l.krasula@gmail.com)
- Joel Jung (joeljung@tencent.com)
- Femi Adeyemi-Ejeye (femi.ae@surrey.ac.uk)
- Denkhan Samilgil (denkhan.samilgil@divitel.com)
- Kjell Brunnström (kjell.brunnstrom@ri.se)
- Steve Göring (steve.goering@tu-ilmenau.de TU Ilmenau)
- Alexander Raake (alexander.raake@tu-ilmenau.de)
- Rakesh Rao Ramachandra Rao (rakesh-rao.ramachandra-rao@tu-ilmenau.de)
- Glenn Van Wallendael (Glenn.VanWallendael@UGent.be)
- Marta Orduna (moc@gti.ssr.upm.es)
- Pramit Mazumdar (pramit.mazumdar@iiitvadodara.ac.in)
- Jing Li (jing.li.univ@gmail.com)
- Saman Zadtootaghaj (Saman.Zadtootaghaj@dolby.com)
-

Introduction

During the development of new video coding standards, peak-signal-to-noise-ratio (PSNR) has been traditionally used as the main objective metric to determine which new coding tools to be adopted. It has been further used to establish the bitrate savings that a new coding standard offers over its predecessor through the employment of the

so-called “BD-rate” metric [reference to Bjontegaard Delta original contribution to MPEG] that still relies on PSNR for measuring quality.
Although this choice was fully justified for the first image/video coding standards - JPEG (1992), MPEG1 (1994), MPEG2 (1996), JPEG2000 and even H.264/AVC (2004) - since there was simply no other alternative at that time, its continuing use for the development of H.265/HEVC (2013), VP9 (2013), AV1 (2018) and most recently EVC and VVC (2020) [need references for all these standards] is questionable, given the rapid and continuous evolution of more perceptual image/video objective quality metrics, such as SSIM (2004), MS-SSIM (2004) and VMAF (2015) [references].

This project attempts to offer some guidance to the video coding community, including standards setting organization (SSOs), on how to better utilize existing objective video quality metrics to better capture the improvements offered by video coding tools.

Scope and goal of the project

1. Address video compression and scaling impairments only
2. First priority should be on full-reference (pixel) objective metrics
3. Examine applicability of no-reference objective metrics
    a. Examples of existing non-ref metrics?
4. Use existing "state-of-the-art" such Full Reference (FR) metrics
    a. Include all metrics currently calculated by the VMAF library: PSNR, SSIM, MS-SSIM, VMAF. Complexity of running the full VMAF library is as follows:
        i. <1 sec/frame, UHD>
    b. Additional metrics - please include computational complexity of each proposed program:
        i. ITU recommendation <place holder>
        ii. fume (full reference pixel only) (non ITU model): https://github.com/Telecommunication-Telemedia-Assessment/pixel models (approx 10s/frame, depending on your PC)
    c. Even though closed-source metrics could perform better, the standardization community is looking for open-source solutions
5. Offer temporal aggregation methods of image quality metrics (such as PSNR and SSIM) into video quality metrics
6. Present statistical analysis of existing subjective datasets, constraining them to compression and scaling artifacts

    a. The databases that can not be shared need to be processed by the owners through running objective metrics locally
    b. Databases that can be shared should be collected and made available to all labs for calculating objective metrics
    c. Labs with subjective data, alongside SRC and PVS, can list their public datasets in the following list:
        i. AVT-VQDB-UHD-1 (TU Ilmenau can also perform additional objective metric calculation for our dataset, we have already VMAF and co included-- probably an update is required)
        ii. <<>>
    d. Labs that have private datasets, can list their objective metric results, with subjective scores (MOS) and bitrates, without SRC and PVS, in the following list
        i. <>
7. Obtain reference implementations of such FR metrics, in order to avoid confusion that happens often, when researchers quote these metrics; for example, there are at least 4 different ways to aggregate PSNR scores, and 3 popular implementations of SSIM
8. Highlight differences among objective metrics and use-cases: for example, in case of very small differences, which metric is more sensitive? Which quality range is better served by what metric?
9. Offer standard logistic mappings of objective metrics to a normalized linear scale (0-100 ?)

Tentative deadline and milestones

- Confirm interest from parties that expressed such during Dec. 2020 VQEG meeting (PIC: Ioannis K, Due date: Mar. 31, 2021)
- Establish a group reflector/mailing list, means of collaboration for the group (PIC: TBD, Due date: Apr. 15, 2021)
- Finalize this document within the IGVQM group, after 1-2 meeting (PIC: TBD, Apr. 30, 2021)
- Send liaison statement to MPEG/ITU-T SG16 and AOM-CWG and solicit interest/feedback
- Collect objective metrics, with corresponding reference implementations from “proponents” that can be used as hypothetical reference circuits (HRCs)

- Collect subjective databases, ideally complete with sources (SRCs), processed video sequences (PVSs), set of opinion scores and details about the methodology used to collect them
- Perform rigorous statistical analysis of opinion scores, utilizing best practices, as captured by the whole VQEG community and in particular from the Statistical Analysis Methods (SAM) group. Ideally, there would be a minimum of 2 groups performing analyses, so there is cross-checking of results.
- Offer linearized versions of proposed metrics, through standard logistic function, such that scores can be normalized to the same, [0-100] linear scale
- Collect all results in a comprehensive report and establish conclusions that include but are not limited to:
    - Identifying which variants of existing metrics are performing best for measuring compression and scaling artifacts (e.g.: which weighting of PSNR-Y/PSNR-Cb/PSNR-Cb is optimal)
    - Propose appropriate temporal aggregation strategies for image quality metrics
    - Test fusion of existing metrics using the VMAF, or other machine-learning based framework, to propose suitable weights to fuse existing metrics
    - Analyze performance of objective metrics in different use-cases, for example low-quality range vs. high-quality range
    - Calculate and establish confidence intervals for the various metrics
- Communicate and solicit feedback from MPEG/AOM on the draft versions of this comprehensive report
- Present the report during VQEG “face-to-face” meeting and ratify after integrating feedback
- Publish the final report on VQEG’s website and submit it to suitable journal/conference

APPENDIX

A. RAW MATERIAL taken from the initial email by IoK and VQEG Dec. 2020 presentation

==========

There is a need from the video compression community for an "implementer's guide" when it comes to objective video quality metrics.

More specifically, as you may know, both MPEG/ITU standardization and AOM (the alliance for open media) is heavily relying on using PSNR to guide their decisions in choosing coding tools for new video codecs. This has been the practice for more than 25 years - and I believe it should be updated in light of the more powerful objective video quality metrics that better correlate with subjective video quality. Since I'm personally involved with these standardization activities

The scope and goal of this contribution/recommendation, the way I see it, is the following:

1) Address video compression and scaling impairments
2) Be constrained to full-reference (pixel) objective metrics
3) List "state-of-the-art" such FR metrics (e.g. PSNR, SSIM, VMAF, etc. - please note that, even though proprietary metrics could perform better, the standardization community is looking for open-source solutions)
4) Offer temporal aggregation methods of frame-level metrics (such as PSNR and SSIM)
5) Present statistical analysis of existing subjective datasets, constraining them to compression and scaling artifacts
6) Obtain reference implementations of such FR metrics, in order to avoid confusion that happens often, when researchers quote these metrics; for example, there are at least 4 different ways to aggregate PSNR scores, and 3 popular implementations of SSIM
7) Highlight differences among objective metrics and use-cases: for example, in case of very small differences, which metric is more sensitive? Which quality range is better served by what metric?
8) Offer standard logistic mappings of objective metrics to a normalized linear scale (0-100 ?)
9) Anything else ?

Let me know what you think - I can take the lead in drafting a document but I foresee that there is need for mostly data coming from various VQEG groups and independent groups that are routinely conducting video subjective experiments.

B. Publicly Available Datasets

From Margaret Pinson:
This web page has a list of datasets that are publicly available and suitable for developing and evaluating NR metrics:
https://github.com/NTIA/NRMetricFramework/blob/master/documentation/SubjectiveDatasets.md
Most of these datasets contain camera impairments, so there is no “source” version of the image or video.

The following video quality datasets are available on CDVL and are suitable for evaluating FR and RR metrics: VQEG datasets FRTV Phase I (2 older datasets), VQEG HDTV (5 datasets), and maybe VQEG MM2 (many ratings but limited content). Log onto CDVL, select advanced search, and from the dataset pull-down, select “VQEG Subjective Tests.”

The following video quality datasets are available on CDVL and may be suitable for evaluating FR and RR metrics: T1A1 (very old), ITS4S, BVI-HD, AGH/NTIA (maybe, bit-rates not available), UnB-AVQ-2013, UnB-AVQ-2018, UnB-3D, AGH-Dolby-NTIA, ITS 2010, ITS AV-Sync 2010, and UnB Varium. Log onto CDVL, select advanced search, and from the dataset pull-down, select “Subjective Tests.” Someone will need to examine the experiment designs and consider whether these datasets are suitable for this project.

The University of Texas at Austin LIVE team has made datasets available. See https://live.ece.utexas.edu/research/Quality/index.htm Some of these datasets may be suitable.