 ISO/IEC JTC 1/SC 29/AG 5 N00058

**ISO/IEC JTC 1/SC 29/AG 5**

**MPEG Visual Quality Assessment**

**Convenorship: DE**

**Document type:** Output Document

**Title:** **Learning-based quality metrics and 2D video databases under study (v2)**

**Status:** Approved

**Date of document:** 2022-06-15

**Source:** ISO/IEC JTC 1/SC 29/AG 5

**Expected action:** Information

**Action due date:** None

**No. of pages:** 5 (without cover page)

**Email of Convenor:** wien@lfb.rwth-aachen.de

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29ag5>

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/AG 5**

**MPEG VISUAL QUALITY ASSESSMENT**

**ISO/IEC JTC 1/SC 29/AG 5 N00058**

**Online – April 2022**

|  |  |
| --- | --- |
| **Title** | Learning-based quality metrics and 2D video databases under study (v2) |
| **Source** | AG 5 MPEG Visual Quality Assessment |
| **Status** | Approved |
| **Serial Number** | 21602 |
| **Authors** | Y. Ye (Alibaba), J. Jung (Tencent), M. Wien (RWTH Aachen University) |

# Introduction

This document provides a summary of learning-based quality metrics and video databases that are currently under study by AG 5 experts, specifically in the AHG on learning-based quality metrics for 2D video. This AHG was established at the 3rd AG 5 meeting in April 2021, and has been mainly focused on studying the potential suitability of a few well-known learning-based quality metrics for use in video coding and video codec development.

Studying the correlation between objective quality metrics and mean opinion scores (MOS) from subjective evaluation involves extensive steps, including:

1. gaining theoretical/mathematical understanding of learning-based quality assessment methodology,
2. selecting well-founded learning-based quality metrics for study,
3. selecting and gathering video datasets with sufficient MOS data that are available for use by MPEG and by JVET,
4. quantifying correlation between objective and subjective scores with sufficient granularity so as to provide meaningful guidance for video coding tool evaluation, and
5. covering different types of video coding technologies including those based on traditional methods and the newly-emerged neural network-based video coding methods.

By the conclusion of the 7th AG 5 meeting in April 2022, correlation studies are still ongoing, and no concrete conclusion has been reached in terms of which learning-based metric(s) may be promising for use in future video coding tool development in MPEG and JVET. This document provides an inventory of learning-based quality metrics and video datasets, as well as a summary of the correlation studies that has been performed so far.

# List of learning-based quality metrics under study

In the development of generations of widely-adopted video coding standards such as H.264/AVC, H.265/HEVC and the latest H.266/VVC, objective metrics such as the Peak Signal to Noise Ratio (PSNR) are often used to measure the distortion between the reconstructed video and the uncompressed original video. Coding tools are evaluated based on their rate-distortion performance as well as their computational and implementation complexity during decision-making. After a standard is finalized, formal subjective testing is performed to verify the overall compression capability of the newly developed standard.

Although PSNR provides a convenient way to assess distortion, it is often found to not align very well with perceived visual quality [1]. Other objective quality metrics have been developed over the years to align objective quality evaluation more closely with the result of subjective quality assessment. Two of the most well-known and representative among these is the structural similarity (SSIM) index [2] and a widely-used variant of it called multi-scale structural similarity (MS-SSIM), which considers the sensitivity of human visual system (HVS) with respect to structural information.

Recent advancement of deep learning has achieved substantial breakthroughs in numerous visual computing tasks, and learning-based quality assessment has received much attention. The following is a short list of industry-adopted and/or highly-cited learning-based quality metrics with open-source implementations:

* **Video Multimethod Fusion Approach (VMAF):** Netﬂix-developed VMAF is an open-source, learning-based full-reference video quality assessment model. It uses a Support Vector Machine (SVM) regressor to combine three elementary video quality features to achieve high correlation with subjective MOS data [3]. The source implementation of VMAF is available at <https://github.com/Netflix/vmaf>.
* **Learned Perceptual Image Patch Similarity (LPIPS)** and **Deep Image Structure and Texture Similarity (DISTS)**: LPIPS and DISTS extract deep features stack from layers of pre-trained network such as ImageNet classification tasks. Authors have shown superior performance in terms of correlation with subjective MOS data from public datasets [4][5]. Implementation of LPIPS and DISTS is available at <https://github.com/dingkeyan93/IQA-optimization>.

# List of 2D video databases under study

## Public datasets

Input contribution m57121 (July 2021) documents a comprehensive list of datasets with MOS data that could be used as ground truth for correlation study when evaluating objective quality metrics. These include datasets widely used in publications and datasets from other standards bodies such as the video quality experts group (VQEG). Based on the initial list, the following three datasets have been selected for AHG study considering many factors such as content quality in the dataset, whether distortion applied to degraded video is due to compression, video codecs that are used and associated coding configurations, and type of MOS scoring method:

1. **Netflix Public dataset [6]:** Netflix dataset contains 34 video clips from various Netflix films and natural scenes, which have different resolutions (from 288P to 1080P). The distorted video sequences are compressed by H.264/AVC with different bitrates (from 0.37Mbps to 20Mbps).
2. **BVI-CC dataset [7]:** BVI-CC dataset contains three groups including UHD, HD and HD-DO, which are selected from Harmonic, BVI-Texture and JVET CTC datasets with natural scenes. The reference sequences from group HD and HD-DO are downscaled from group UHD. The distorted video sequences from UHD and HD groups are compressed by AV1, HM and VTM with 4 rate points for each reference sequence. Group HD-DO contains videos compressed for three different resolutions (1920×1080, 1280×720, and 960×544), the reconstruction of which are up-sampled to HD resolution using Lanczos-3 filter. These distorted video sequences are compressed by AV1 and HM with 5 rate points. The subjective test follows the Double Stimulus Continuous Quality Scale (DSCQS) methodology [9], and Difference Mean Opinion Scores (DMOS) were then obtained for each trial by taking the mean of the difference scores among participants.
3. **TUM 1080p25 dataset [8]:** TUM 1080p25 dataset consists of four different HDTV 1080p25 video sequences from the SVT test set with natural scenes. The distorted video sequences are compressed by H.264/AVC and Dirac with different bitrates (from 5.4 Mbit/s to 30Mbit/s). Its subjective test follows the Double Stimulus Unknown Reference (DSUR) [10] method, and the MOS is obtained by averaging all valid votes for each test case, ranging from 0 to 10.

## JVET remote expert viewing datasets

Remote expert viewing tests were conducted during several JVET meetings for the exploration activity on neural network-based coding tools. The testing procedure followed the AG 5 guidelines for remote experts viewing [11], providing an A/B comparison of sequences under test. The viewing compared the reconstruction quality of several proposals under test to the reconstruction quality of the VTM anchor. The votes of the test subjects, using this Comparison Category Rating (CCR) method [12], are mapped to Comparison Mean Opinion Scores (CMOS), with numbers as follows:

|  |
| --- |
| +3 : “Proposal much better than anchor”  +1 : “Proposal better than anchor”  -1 : “Anchor better than proposal”  -3 : “Anchor much better than proposal” |

From these values, the CMOS values were calculated. The scores of the four participants with the lowest Pearson correlation were removed from the evaluation. The proposals under test, test material used, and MOS data are documented in JVET-U0142, JVET-V0173, JVET-W0186, JVET-X0209, JVET-Y0212, and JVET-Z0053 and the test sequences can be retrieved from JVET ftp site (ftp.ient.rwth-aachen.de).

# Correlation study

## Public dataset correlation study

In contribution m57993 (October 2021), a correlation study between the subjective MOS scores as ground truth from the three public datasets in section 3.1 and six objective quality metrics (including three traditional metrics and three learning-based metrics) is conducted. The correlation study uses widely-adopted criteria including Pearson linear correlation coefficient (PLCC), the Spearman rank correlation coefficient (SRCC), and the Kendall rank correlation coefficient (KRCC).

Table 1. Average correlation performance comparison of traditional and learning based quality metrics across different datasets, sorted according to ranking.

|  |  |  |  |
| --- | --- | --- | --- |
| Method | PLCC(Rank) | SRCC(Rank) | KRCC(Rank) |
| VMAF | 0.859(#1) | 0.852(#1) | 0.666(#1) |
| DISTS | 0.814(#2) | 0.803(#2) | 0.619(#2) |
| MS-SSIM | 0.735(#3) | 0.764(#3) | 0.580(#3) |
| SSIM | 0.723(#4) | 0.706(#4) | 0.529(#4) |
| LPIPS | 0.634(#6) | 0.641(#5) | 0.477(#5) |
| PSNR | 0.654(#5) | 0.632(#6) | 0.466(#6) |

Table 1 shows the correlation scores averaged over all three datasets and sorted from high to low for the six representative quality metrics. The following observations can be made:

1. Two learning-based quality metrics, VMAF and DISTS, are consistently ranked #1 and #2 among all metrics, and consistently outperform the traditional quality metrics;
2. Among the traditional metrics, the widely used MS-SSIM exhibits the highest correlation with subjective quality (this is consistent with previously published literature);
3. PSNR, the most widely used in MPEG coding standard development, has low correlation with subjective quality (this is consistent with previously published literature), and is far behind top-ranked metrics like VMAF and DISTS in terms of predicting subjective quality;
4. Regardless of different methods of computing correlation (that is, PLCC, SRCC, or KRCC), the ranking of each quality metric is mostly consistent.

Table 1’s results are averaged over all three datasets and include a number of different video codecs, some MPEG and some non-MPEG. The correlation study in m57993 also includes more codec-specific correlation data. Though not included herein for simplicity, very similar observations as the above can be made based on those codec-specific correlation data.

In contribution m59384 (April 2022), 13 different full-reference metrics are evaluated on the most recent dataset only (BVI-CC). It is reported that VMAF and AVQT are the most efficient and robust metrics on this dataset, in the sense that they are consistently providing the highest correlations and lowest RMSE, no matter the tested configuration, as shown in figures 1 and 2. The contribution also confirms the efficiency of DISTS already demonstrated in m57993, and highlights once again the weakness of the PSNR as having very low correlation with subjective results. As additional results, it is shown that the correlation is similar for HD and UHD content (except for some learning-based metrics), that the correlation of several metrics is excellent for VVC (reaching 0.9 PLCC), but none of them is able to assess correctly the quality of AV1, and finally that the metrics are more reliable at low bitrate, and for content with low spatial information and high temporal information.

|  |  |
| --- | --- |
| Inserting image...  Figure 1: correlation between MOS and metrics on HD and UHD content, full reference metrics | Inserting image...  Figure 2: RMSE between MOS and metrics on HD and UHD content, full reference metrics |

## JVET dataset correlation study

Correlation study with JVET remote expert viewing has been performed in several contributions.

Input contribution m59045 (Jan 2022) applies quantitative correlation study (using PLCC, SRCC, and KRCC) between CMOS and all the quality metrics in Table 1 except LPIPS, and finds that the linear correlation on the JVET dataset appears to be quite poor for all of the five quality metrics, regardless of whether they are learning based or not. While some possible explanations exist, ranging from the relatively small size of the JVET test datasets, to how remote viewing tests were conducted, and further to how the correlation was calculated, no concrete reasons could be verified/confirmed, and future investigation is needed.

In contribution m58816 (Jan 2022), the ability of fifteen objective metrics to answer to the question “is a proposal better than an anchor?” is evaluated. It is explained that no quantitative correlation is applied (PLCC, SRCC, KRCC): the non-linearity of the JVET scale would necessarily yield to poor correlations, not representative of the metrics’ performances. On the contrary, a ROC analysis is considered, adapted to the CCR methodology used to produce the JVET dataset.

The contribution m59383 (April 2022) updates m58816 by:

* considering the scores produced at the JVET January meeting,
* considering subjective scores bias resulting from the forced-choice scale,
* considering confidence intervals,
* adding two more metrics (DISTS and 3SSIM), and preliminary results by splitting the results per categories of proposals.

It is reported that the correct decision rate is around 84% for SSIM, VMAF and LPIPS (with VGG network), as shown in figure 3, for the cases where the proposal and the anchor are considered as different, according to subjective scores (CMOS). These three metrics are on average the most efficient, when considering different conditions and situations. In addition, the PSNR is consistently less reliable than VMAF in all tested configurations.

Chart, bar chart

Description automatically generated

Figure 3: correct decision rate to the question “is a proposal better than an anchor”, for a set of objective quality metrics, with constraint on the CMOS values

# References

1. W. Lin, D. Li, and P. Xue, “Discriminative analysis of pixel difference towards picture quality prediction,” in Proceedings 2003 International Conference on Image Processing, vol. 3. IEEE, 2003, pp. III–193.
2. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.
3. J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C. J. Kuo, “A fusion-based video quality assessment (fvqa) index,” in Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Paciﬁc. IEEE, 2014, pp. 1–5.
4. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp.586–595.
5. K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.
6. Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, “Toward a practical perceptual video quality metric,” The Netflix Tech Blog, vol. 6, 2016.
7. A. V. Katsenou, F. Zhang, M. Afonso, and D. R. Bull, “A subjective comparison of AV1 and HEVC for adaptive video streaming,” in Proc. IEEE Int Conf. on Image Processing, 2019, pp. 4145–4149.
8. C. Keimel, J. Habigt, T. Habigt, M. Rothbucher, and K. Diepold, “Visual quality of current coding technologies at high definition IPTV bitrates,” in Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on, 2010, pp.390 –393
9. Recommendation ITU-R BT.500-12, Methodology for the subjective assessment of the quality of television pictures, 2012.
10. V. Baroncini, “New tendencies in subjective video quality evaluation,” IEICE Transaction Fundamentals, vol. E89-A, no. 11, pp. 2933–2937, 2006.
11. J. Jung, M. Wien, V. Baroncini, Guidelines for remote experts viewing sessions, ISO/IEC JTC 1/SC 29/AG 5 N00040 Online, October 2021.
12. Recommendation ITU-T P.800, Methods for subjective determination of transmission quality, 1996.