

Applying Objective Quality Metrics to Video-Codec Comparisons: Choosing the Best Metric for Subjective Quality Estimation

Anastasia Antsiferova¹, Alexander Yakovenko¹, Nickolay Safonov¹,
Dmitriy Kulikov^{1,2}, Alexander Gushin¹ and Dmitriy Vatolin¹

¹Lomonosov Moscow State University, Leninskiye Gory, 1, Moscow, 119991, Russia

²Dubna State University, Universitetskaya, 19, Dubna, 141982, Russia

Abstract

Quality assessment is essential to creating and comparing video compression algorithms. Despite the development of many new quality-assessment methods, well-known and generally accepted codecs comparisons mainly employ classical methods such as PSNR, SSIM, and VMAF. These methods have different variations: temporal pooling techniques, color-component summations and versions. In this paper, we present comparison results for generally accepted video-quality metrics to determine which ones are most relevant to video codecs comparisons. For evaluation we used videos compressed by codecs of different standards at three bitrates, and subjective scores were collected for these videos. Evaluation dataset consists of 789 encoded streams and 320294 subjective scores. VMAF calculated for all Y, U, V color spaced showed the best correlation with subjective quality, and we also showed that the usage of smaller weighting coefficients for U and V components leads to a better correlation with subjective quality.

Keywords

video quality rating, comparison of metrics, video codecs comparison, PSNR, SSIM, VMAF

1. Introduction

A Cisco forecast [1] predicts that by 2022, 79% of the world's Internet traffic will be video. To reduce the cost of video storage and the burden on data-transmission channels, creation and improvement of video-compression algorithms are under way. Video-quality measurement is crucial in this area. The number of studies, publications, and grants allocated to the development of new quality metrics is growing yearly. One reason why new metrics are seldom used is that their accuracy is unreproducible on large sets of real data. Generally accepted comparisons of compression algorithms, therefore, still employ classical methods: PSNR, SSIM, and the new VMAF, which has gained popularity. For example, to demonstrate the effectiveness of a new

GraphiCon 2021: 31st International Conference on Computer Graphics and Vision, September 27–30, 2021, Nizhny Novgorod, Russia

✉ aantsiferova.graphics.cs.msu.ru (A. Antsiferova); alexander.yakovenko@graphics.cs.msu.ru (A. Yakovenko); nikolay.safonov@graphics.cs.msu.ru (N. Safonov); dkulikov@graphics.cs.msu.ru (D. Kulikov); alexander.gushchin@graphics.cs.msu.ru (A. Gushin); dmitriy@graphics.cs.msu.ru (D. Vatolin)

🆔 0000-0002-1272-5135 (A. Antsiferova); 0000-0003-3105-512X (A. Yakovenko); 0000-0001-9950-2403 (N. Safonov); 0000-0003-1264-2118 (D. Kulikov); 0000-0002-4055-7394 (A. Gushin); 0000-0002-8893-9340 (D. Vatolin)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

coding standard, the pertinent committee objectively tests multiple videos using PSNR [2], followed by subjective testing. Companies involved in developing new video codecs, as well as their customers, have begun using VMAF, which in many studies has shown a high correlation with visual quality [3].

All of the above methods are common ways to increase quality control; for example, owing to its high computation speed, PSNR serves in early development stages, in which thousands of configurations require testing. Intermediate stages employ SSIM and VMAF, since they take longer to compute. Generally accepted comparisons typically use all these metrics—examples include the work of Jan Ozer [4], as well as annual comparisons by Moscow State University [5]. But these metrics allow many configurations that affect the ranking of compression methods. For instance, the PSNR and SSIM calculations can only be performed on the luminance Y channel; another option is summation of the Y, U, and V channels. The sum can also use different coefficients (luminance usually has a larger coefficient than other channels). Currently, owing to a lack of recommended methods for sums and for using certain channels to calculate metrics, interpretation of the comparison results becomes much more complicated: instead of just 3 metrics, 20 or more may be necessary.

Many implementations of modern compression standards have special modes to increase their output's score on popular metrics. For example, x264 and x265 have configuration modes for PSNR and SSIM. The libaom encoder has a VMAF tuning mode. That tuning information is visible, but many commercial solutions contain hidden settings to increase their scores on generally accepted metrics, potentially reducing visual quality. A subjective comparison at Moscow State University showed that the “-tune ssim” setting improves visual quality in addition to SSIM scores. But the video-preprocessing techniques in libaom's “-tune vmaf” setting can substantially reduce visual quality, as [6] demonstrated.

For this paper, we analyzed the correspondence to visual quality of various PSNR, SSIM, and VMAF configurations. We collected a special data set for our analysis, as open data sets with visual-quality ratings contain distortions from just one or two codecs (usually H.264 and H.265). Therefore, our study paid special attention to assembling a set of videos encoded by various implementations of multiple standards. In this way, we obtained videos with representative encoding distortions. This task was under way from 2018 to 2021 through the annual subjective video-codec comparisons of Moscow State University [5].

2. Related work

Multiple studies have compared video-quality metrics. At the same time, each metric's correlation with subjective estimates can vary greatly depending on the data set and the distortion type. For example, the goal of [7] was to show that metrics targeting TV signals with high resolution, high bitrate, and high frame rate (FPS) perform poorly on video with low bitrate, low resolution, and variable FPS. The researchers confirmed this conclusion and showed that the NTIA videoconference model [8] delivers the best accuracy, followed by the NTIA general model [8], Watson's DVQ [9], and VSSIM [10]. The authors of [11] compared nine metrics on three data sets: LIVE, ECVQ, and EVVQ. The latter two were created using the JVT JM v.10.2 encoder (based on H.264/AVC) and XviD v.1.1.0 (an open-source encoder based on the

MPEG-4 Part 2 specification). The authors concluded that the Movie and FMSE metrics perform better than others when assessing all impairments, except when simulating transmission over an IP network. In [12], the researchers compared various metrics using 20 FullHD sequences from popular streaming services. Employing the x265 v2.7 encoder, they compressed these sequences to 10 quality levels defined by gradually increasing bitrate and resolution. Their study employed five metrics to assess the quality of the resulting videos. They concluded that the metrics correlate much better in the SD range than in the HD range, which has noticeably fewer compression artifacts. In this comparison, VMAF demonstrated better correlation than PSNR, SSIM, MS-SSIM, and VIF.

Thus, finding a suitable quality metric that maximally correlates with visual assessment is essential. The existing works on the correlation of video-quality metrics examine only a few video codecs (mainly open implementations of H.264 and H.265). Therefore, evaluating a metric's relevance to a wide variety of videos containing many types of distortions remains important. Numerous data sets are useful for comparing algorithm performance, the most popular being Live-VQA [13] and Live-VQC [14]. Their biggest drawback when attempting to identify the best compression-quality metrics is that they contain few compression artifacts. Also, many new metrics that employ machine learning have been trained on these data sets. These factors call into question the applicability of such data sets to objective metric comparisons and make beneficial a study that uses an independent data set with a representative spectrum of compression artifacts.

In addition to the variety of metrics, there are several ways to calculate them: using RGB or YUV color models, using only the luminance component, or using all color panes with different summation coefficients. The YUV color space is a common choice for image- and video-quality measurement [15]. It enables quality measurement using only the Y (luminance) space, which provides more visual information about an image. The U and V spaces have less impact; for some metrics they improve overall correlation but require additional computation time. In this paper we show the efficiency of summing different Y, U, and V components for different metrics.

3. Data collection for evaluating video-quality-assessment algorithms

To analyze the relevance of quality metrics to codec comparisons, we collected a special data set that includes video sequences and subjective scores. The subjective comparisons were performed independently using different videos and a different encoders sets. Each one assessed FullHD videos with different spatial and temporal complexities, which affect compression quality and performance. We made our selection from a pool of more than 18,000 open-source clips with high bitrate after analyzing more than five million source videos from the Vimeo website. Our choice employed clustering in terms of space-time complexity. A description of the video-selection method appears in [16]. The resulting video data sets for subjective assessments are called CC-2018, CC-2019, CC-2020, and UGC-2020; the numbers indicate the year each one was created. Each data set is available by request from its associated codec-comparison project page [5].

We obtained a representative set of coding artifacts using different video-codec standards: 11

Table 1

Data sets for analyzing video-quality metrics. CC-2018, CC-2019, CC-2020 and UGC-2020 are video sets used for subjective evaluation in MSU Codecs Comparisons in 2018, 2019 and 2020.

Video dataset	Number of codecs	Number of test videos	Number of encoded streams	Number of responses
CC-2018	10	5	150	22542
CC-2019	11	5	165	25784
CC-2020	11	8	264	236736
UGC-2020	7	10	210	35232
Total	39	28	789	320294

H.265/HEVC encoders, five AV1 encoders, two H.264/AVC encoders, and four encoders based on other standards (VVC, VP9, SIF, and xvc). We compressed each video at three target bitrates: 1,000 Kbps, 2,000 Kbps, and 4,000 Kbps. The choice of this range simplifies the subjective-comparison procedure, since the video quality is more difficult to distinguish visually at higher bitrates.

The subjective assessment involved pairwise comparisons using the Subjectify.us platform, which employs a Bradley-Terry model to transform the results of pairwise voting into a score for each video. A detailed description of the method appears on the website. To increase the relevance of the results, each pair of videos received at least 10 responses from participants. The number of subjective ratings per pair depended on the confidence intervals: more responses were received for complex videos as well as videos that were hard to be compared.

Table 1 summarizes the data sets, which the 2018–2020 MSU codec comparisons used for subjective evaluation.

We measured various configurations of PSNR, SSIM, MS-SSIM, VMAF, and NIQE for all encoded videos. Our analysis considered the following versions of the PSNR algorithm:

- PSNR average MSE – when aggregating frame-by-frame scores for the entire video, we first calculated the arithmetic mean for the MSE and then the logarithm.

$$\text{PSNR}_{\text{avg.MSE}}(V, \hat{V}) = 10 \log_{10} \frac{\text{MAX}_I^2}{\frac{1}{n} \sum_{i=1}^n \text{MSE}(V_{(i)}, \hat{V}_{(i)})}$$

- PSNR average log – when aggregating frame-by-frame scores for the entire video, we calculate the PSNR for each frame and then the arithmetic mean for all frames.

$$\text{PSNR}_{\text{avg.log}}(V, \hat{V}) = \frac{1}{n} \sum_{i=1}^n 10 \log_{10} \frac{\text{MAX}_I^2}{\text{MSE}(V_{(i)}, \hat{V}_{(i)})}$$

In addition, we considered different versions of the VMAF metric:

- VMAF 0.6.1, VMAF 0.6.2, and VMAF 0.6.3.
- VMAF 0.6.1 NEG (“no enhancement gain”), which is less prone to artificial increases through preprocessing.

- The “Phone” variant of the above four, as well as a variant that handles 4K video using the VMAF 0.6.1 model.

For each reference method, our analysis considered the following options for calculating the color components: we evaluated Y metrics only for the luminance channel, we evaluated YUV4:1:1 metrics independently for the three components and averaged the result as $4*Y+U+V$, and we applied similar methods for YUV6:1:1, YUV8:1:1, YUV10:1:1, and the rarely used YUV1:1:1 as well as YUV2:1:1. Our calculations of the metric values employed the MSU VQMT version 12.1 [17].

4. Results

Because we conducted a separate subjective comparison of the videos for each year, we had to obtain for every metric an overall correlation across the entire data set. For each metric, our approach applied the Fisher z-transform to all correlations coefficients calculated on individual video sequences and used the resulting values to calculate the weighted mean and confidence interval, with the weights proportional to the number of distortions. We determined the final correlations using the inverse transformation [18].

Fig. 1 and Fig. 2 show the Spearman and Pearson correlations. The colors indicate different groups of metrics. The graphs reveal that VMAF variations have the highest correlation with subjective-quality scores, and the differences between them are insignificant. Next are VMAF NEG and MS-SSIM, also with nearly equal correlation. PSNR variants have the lowest correlation among full-reference metrics.

4.1. Comparison of YUV summations

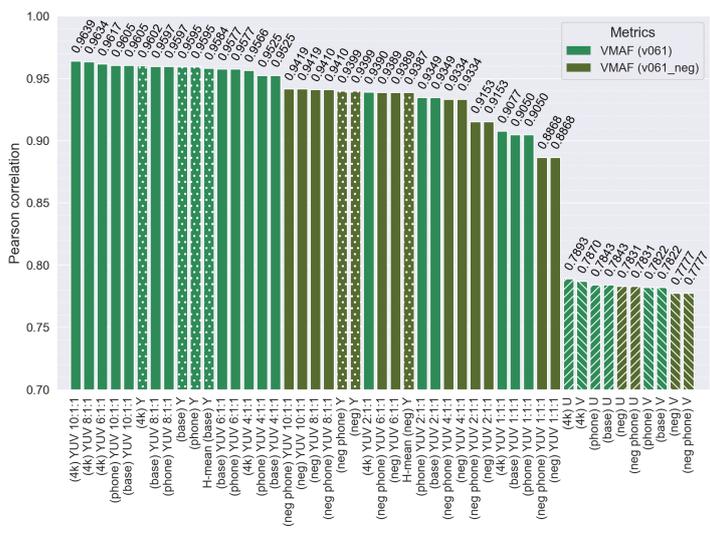
In all cases, YUV1:1:1 and YUV2:1:1 show worse results than the other channel-summing methods. Different versions of some metrics (for example, MS-SSIM) exhibit nearly identical results, but no one YUV summation is best for all metrics.

Table 2 shows the best options for summing the components of different metrics. Some options have almost identical correlation; the table separates them by commas. For VMAF, the best YUV-summation coefficients are 6:1:1, 8:1:1, and 10:1:1 for PSNR, and 6:1:1 and 4:1:1 for SSIM. For MS-SSIM, nearly all methods have the same accuracy.

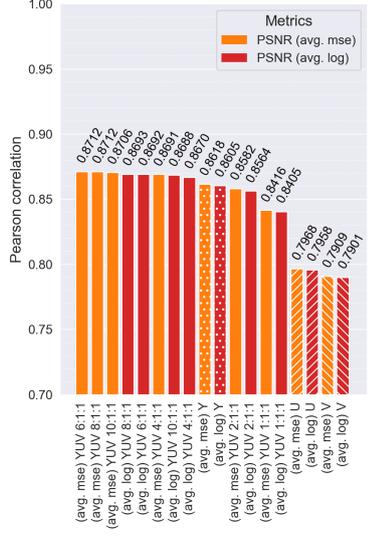
Fig. 3a and Fig. 3b show overall ranking of metrics with best summations only.

4.2. Comparison of VMAF and VMAF NEG

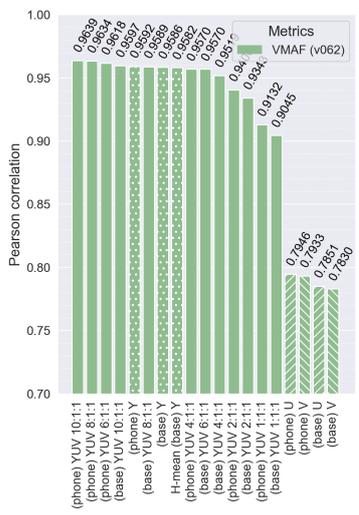
VMAF NEG showed a lesser correlation with visual quality than VMAF did (Fig. 4). Its developer, Netflix, recommends using VMAF NEG when comparing video codecs, as it helps prevent cheating and artificial metric increases through video preprocessing. When comparing different versions of the same algorithm, however, as well as when algorithms are incapable of artificially increasing VMAF, the classic VMAF yields a more accurate visual-quality estimate.



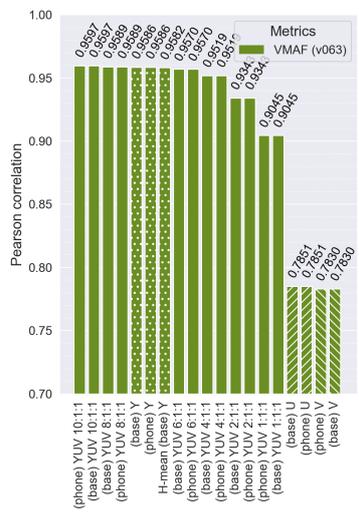
(a) VMAF v0.6.1 including VMAF NEG



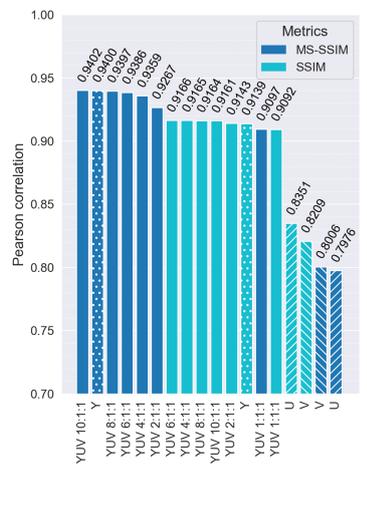
(b) PSNR



(c) VMAF v0.6.2



(d) VMAF v0.6.3



(e) SSIM and MS-SSIM

Figure 1: Pearson correlation between objective-metric scores and the visual-assessment rankings. Comparison of various metrics using the YUV summation technique. The notation (e.g., 4:1:1) indicates the coefficients that are proportional to the metric-value weights for the Y, U, and V color components.

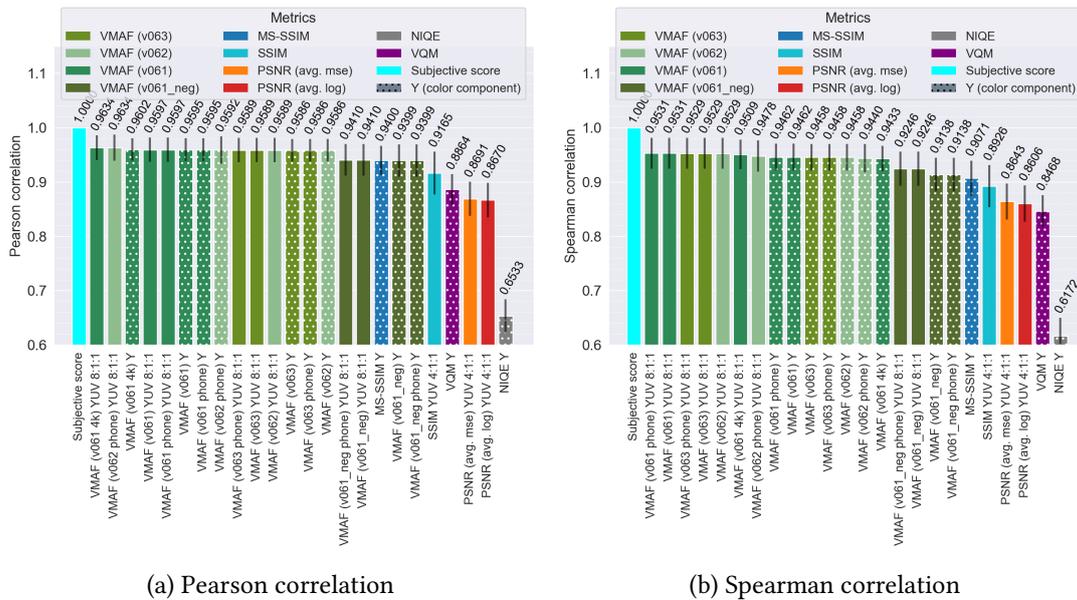
4.3. Comparison of SSIM/MS-SSIM, PSNR average MSE, and PSNR average log

MS-SSIM correlates better with visual quality compared with classic SSIM. Different PSNR variants correlate worse with visual quality than SSIM and MS-SSIM do, but PSNR average MSE is slightly better than PSNR average log (Fig. 2e and Fig. 2b).

Table 2

Y, U, and V summation methods that correlate best with visual quality for different metrics.

Metric	Best YUV Summing
VMAF 0.6.1	
VMAF 0.6.2	
VMAF 0.6.3	
VMAF 0.6.1 phone	
VMAF 0.6.1 neg	8:1:1, 10:1:1, 6:1:1
VMAF 0.6.1 neg phone	
VMAF 0.6.1 4K	
VMAF 0.6.2 phone	
VMAF 0.6.3 phone	
PSNR avg. MSE	6:1:1, 4:1:1
PSNR avg. log	
SSIM	No significant difference between summing methods
MS-SSIM	Y (only luma component), 10:1:1, 6:1:1, 8:1:1

**Figure 3:** Pearson correlation between objective-metric scores and rankings based on visual assessment.

for other standards. This result may be a consequence of recent encoders employing neural networks, which restore object boundaries and thus cause pixel-by-pixel similarity violations—a characteristic that PSNR penalizes. Multiscale metrics (VMAF and MS-SSIM) yielded the best results for such videos. The most stable values are for metrics that evaluate H.264/AVC-encoded streams. The correlations for all metrics, when applied to these videos, exceed 0.94.

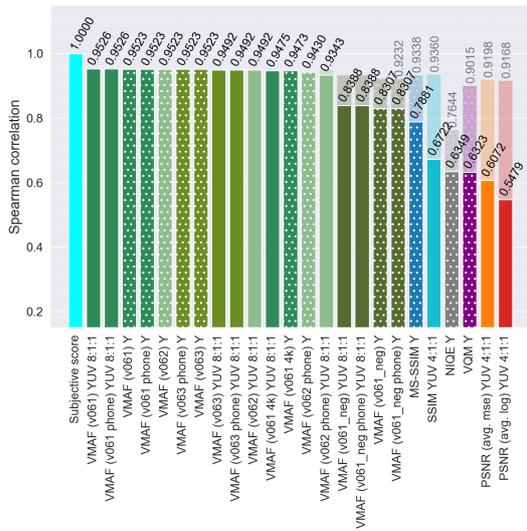


Figure 4: Comparison of metric correlations for video streams encoded by AV1 (saturated colors) and a set of video streams encoded with H.265 encoders (semi-transparent columns). PSNR has low relevance for analyzing the quality of AV1 video streams.

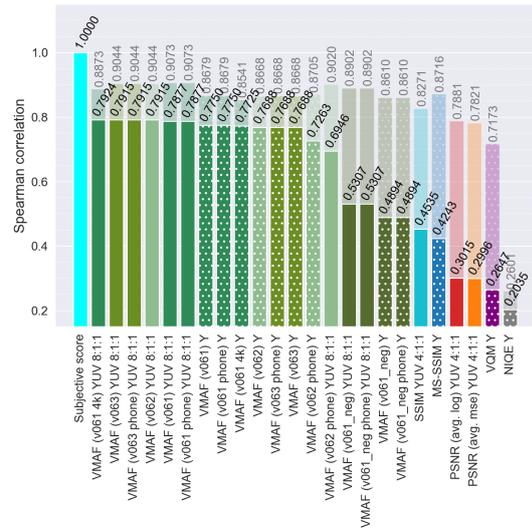


Figure 5: Comparison of metric correlations for video streams encoded with low bitrate (semi-transparent colors) and high bitrate (saturated colors). Video streams with low bitrate are easier to compare visually, and the correlation between metrics and visual scores is higher.

4.5. Metrics comparison for different bitrates

Fig. 5 shows that the relevance of metrics for high-bitrate encoded videos is less than for low-bitrate videos. This result may be due to the difficulty of visually ranking good-quality videos, whereas ranking low-bitrate videos that contain artifacts is easy. We can thus conclude that VMAF reflects the visual perception of artifacts much better than do other metrics for high bitrates.

4.6. Metric comparison for different videos

Fig. 6 shows the space-time complexity distribution and the correlation between PSNR and VMAF for different videos. For some videos that exhibit low spatial and temporal complexity, PSNR and SSIM differ greatly from VMAF. This difference may owe to other factors, so tracking several metrics when measuring the performance on individual videos is better.

5. Conclusion

This article describes the results of comparing different versions of popular objective methods with subjective quality rankings. Our analysis used numerous compression algorithms and revealed the best variants for video-codec comparisons. We used a large-scale data set containing 789 encoded videos distorted by 39 versions of H.264, H.265, AV1, VP9, and other codecs, as well

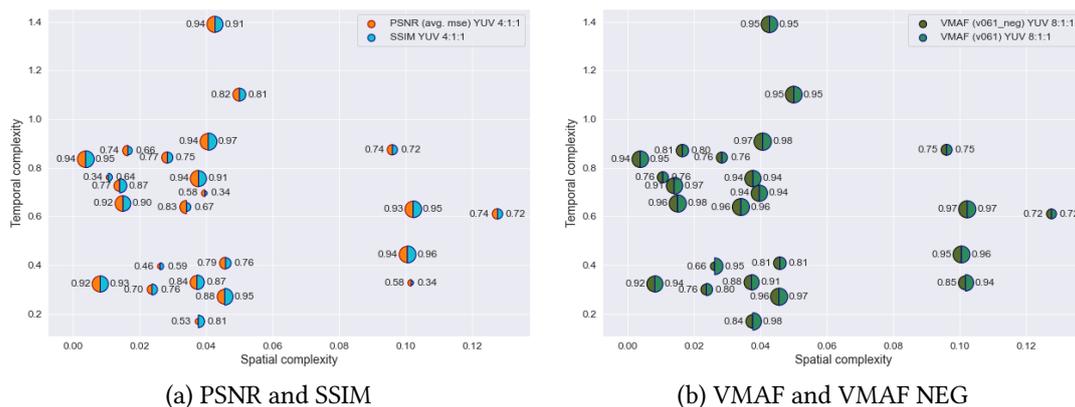


Figure 6: Correlation of different metrics by spatial and temporal complexity.

as three bitrates. We conducted a visual analysis of the resulting sequences using Subjectify.us; several hundred individuals participated. We analyzed many metric versions and modifications (different methods of averaging the values between frames, accounting for color and brightness, and so on).

Analysis of the results led us to the following conclusions:

1. VMAF and its variants exhibited higher correlation with visual quality than other metrics did. Recent research, however, showed that if videos are specially prepared (preprocessed) for this metric [19, 6], visual quality may decline, causing the correlation to become negative. At the same time, for high bitrates, VMAF outperforms the results of other metrics (its correlation is 0.7, versus 0.25–0.45).
2. When calculating metrics for all YUV color planes, different summation methods work best for different metrics:
 - For VMAF, an 8:1:1 ratio provides the best result when summing over Y, U, and V.
 - For VMAF NEG, 6:1:1 is best.
 - For SSIM, 6:1:1.
 - For PSNR (average MSE), 6:1:1.
 - For PSNR (average log), 4:1:1.
3. MS-SSIM showed better results than SSIM.
4. Modifications of PSNR (average log and average MSE) yielded no significant differences.
5. When analyzing AV1 codecs or AV1-encoded videos, no PSNR modifications are justified.
6. Some metrics that have similar average correlations may yield lower-quality results for some videos. Therefore, when comparing the quality of video-encoding or video-processing algorithms, it makes sense to employ several metrics while taking into account their potentially sharp fluctuations for individual outputs.

The above results prove that comparing video-coding algorithms using objective quality metrics is a complex process with many issues and peculiarities. Ignoring these characteristics may lead to results that are unwarranted or, sometimes, that contradict the results of subjective

visual analysis. For that reason, codec-industry professionals recognize only well-known codec comparisons, which should be carried out in laboratories by experienced teams in collaboration with codec developers and other industry experts. These comparisons employ a correct methodological basis and empirical confirmation of their various facets (e.g., choice of objective quality metrics, modifications to those metrics, and methods of averaging color components). Otherwise, given a certain selection of video data, metrics, and metric parameters, a careless or unsuspecting researcher can easily obtain the desired comparison result rather than an objective one.

Acknowledgments

This work is partially supported by the Russian Foundation for Basic Research under Grant 19-01-00785a. Anastasia Antsiferova was supported by the Fellowship from Non-commercial Foundation for the Advancement of Science and Education INTELLECT. Special thanks go to the Graphics and Media Lab at Moscow State University for providing valuable advice and support for our projects.

References

- [1] Cisco vni report 2017-2022, 2018. URL: <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html>.
- [2] J. Boyce, K. Suehring, X. Li, V. Seregin, Jvet-j1010: Jvet common test conditions and software reference configurations, 2018.
- [3] R. Rassool, Vmaf reproducibility: Validating a perceptual practical video quality metric, in: 2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2017, pp. 1–2. doi:10.1109/BMSB.2017.7986143.
- [4] J. Ozer, Av1 has arrived: Comparing codecs from aomedia, visionular, and intel/netflix, 2020. URL: <https://www.streamingmedia.com/Articles/Editorial/Featured-Articles/AV1-Has-Arrived-Comparing-Codecs-from-AOMedia-Visionular-and-Intel-Netflix-142941.aspx>.
- [5] Msu video codecs comparisons, n.d. URL: http://compression.ru/video/codec_comparison/index_en.html.
- [6] M. Siniukov, A. Antsiferova, D. Kulikov, D. Vatolin, Hacking vmaf and vmaf neg: metrics vulnerability to different preprocessing, 2021. arXiv:2107.04510.
- [7] M. H. Loke, E. P. Ong, W. Lin, Z. Lu, S. Yao, Comparison of video quality metrics on multimedia videos, in: 2006 International Conference on Image Processing, 2006, pp. 457–460. doi:10.1109/ICIP.2006.312492.
- [8] S. Wolf, M. Pinson, Video quality measurement techniques, National Telecommunications and Information Administration (NTIA) Report (2002).
- [9] F. Xiao, Dct-based video quality evaluation—final project for ee392j, 2000.
- [10] Z. Wang, L. Lu, A. C. Bovik, Video quality assessment based on structural distortion measurement, *Signal Processing: Image Communication* 19 (2004) 121–132. URL: <https://doi.org/10.1016/j.spi.2004.05.001>.

//www.sciencedirect.com/science/article/pii/S0923596503000766. doi:[https://doi.org/10.1016/S0923-5965\(03\)00076-6](https://doi.org/10.1016/S0923-5965(03)00076-6).

- [11] M. Vranješ, S. Rimac-Drlje, K. Grgić, Review of objective video quality metrics and performance comparison using different databases, *Signal Processing: Image Communication* 28 (2013) 1–19. URL: <https://www.sciencedirect.com/science/article/pii/S0923596512001919>. doi:<https://doi.org/10.1016/j.image.2012.10.003>.
- [12] D. Nandakumar, Y. Wu, H. Wei, A. Ten-Ami, On the accuracy of video quality measurement techniques, in: *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–6. doi:10.1109/MMSP.2019.8901796.
- [13] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, L. K. Cormack, Study of subjective and objective quality assessment of video, *IEEE Transactions on Image Processing* 19 (2010) 1427–1441. doi:10.1109/TIP.2010.2042111.
- [14] Z. Sinno, A. C. Bovik, Large-scale study of perceptual video quality, *IEEE Transactions on Image Processing* 28 (2019) 612–627. doi:10.1109/TIP.2018.2869673.
- [15] M. Podpora, G. P. Korbas, A. Kawala-Janik, Yuv vs rgb-choosing a color space for human-machine interaction., in: *FedCSIS (Position Papers)*, 2014, pp. 29–34.
- [16] A. V. Zvezdakova, D. L. Kulikov, S. V. Zvezdakov, D. S. Vatolin, Bsq-rate: a new approach for video-codec performance comparison and drawbacks of current solutions, *Programming and computer software* 46 (2020) 183–194.
- [17] Msu quality measurement tool, n.d. URL: http://compression.ru/video/quality_measure/vqmt_download.html.
- [18] D. M. Corey, W. P. Dunlap, M. J. Burke, Averaging correlations: Expected values and bias in combined pearson rs and fisher’s z transformations, *The Journal of General Psychology* 125 (1998) 245–261. URL: <https://doi.org/10.1080/00221309809595548>. doi:10.1080/00221309809595548. arXiv:<https://doi.org/10.1080/00221309809595548>.
- [19] A. Zvezdakova, S. Zvezdakov, D. Kulikov, D. Vatolin, Hacking vmaf with video color and contrast distortion, in: *CEUR Workshop Proceedings*, 2019, pp. 53–57.