On the accuracy of open video quality metrics for local decision in AV1 video codec

Andreas Pastor, Lukas Krasula, Xiaoqing Zhu, Zhi Li, Patrick Le Callet

ON THE ACCURACY OF OPEN VIDEO QUALITY METRICS FOR LOCAL DECISION IN AV1 VIDEO CODEC

Andréas Pastor[⋆] *Lukáš Krasula*[†] *Xiaoqing Zhu*[†] *Zhi Li*[†] *Patrick Le Callet*[⋆]

[⋆] Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[†]Netflix Inc., Los Gatos, CA, USA

ABSTRACT

VMAF is a popular objective quality metric used for video quality evaluation. The power of VMAF has been demonstrated for a wide variety of video scales and encoding processes. However, its ability to evaluate the quality of small video patches has not yet been tested, despite its importance for encoding algorithms. We applied Maximum Likelihood Difference Scaling (MLDS) methodology to estimate supra-threshold perceptual differences in localized sections in videos, also known as tubes, encoded using AV1. We further used the results to assess the performance of VMAF in this scenario and proposed a recalibration of the algorithm to improve its agreement with the subjective data.

Index Terms—

Difference Scaling, Video Quality, VMAF, tubes, Open Video Codecs

1. INTRODUCTION

Video encoding and research on human perception of distortion attempt to improve the quality of transmitted videos, while keeping file sizes as small as possible. Multiple objective quality metrics exist and work well on a global scale, but their behavior, when applied locally on small video patches, remains largely an open question.

In video coding algorithms like AV1[1], the optimizer selects at the level of a coding unit (CU) between the different coded proposals of a reference block, taking into account the cost (i.e. bit rate) and the distortion introduced. At equal cost, the most faithful proposal will be retained. This fidelity paradigm guides the encoding and we need a powerful metric to assess the perception of these distortions.

Previous research work such as [1, 2, 3] have used supra threshold perceptual scaling to assess image similarity and apply perceptual weights to the three SSIM [4] terms. Another series of works from [5, 6, 7, 8, 9] applied MLDS for video similarity. In [10, 11], they applied MLDS to assess the perception of encoded textures. In recent work [12], MLDS has been compared to other two-alternated force choice (2AFC) subjective methods in a comparative study to boost data collection. Difference scaling showed good performance in the task of estimating perceived distortions of compressed medias, so we chose to use it in this work.

In this study, we present how we collected data for supra-threshold perceptual distortion introduced with AV1 in videos. In section 2, we detail the creation of tube-contents and our strategy to select interesting ones for a subjective experiment using MLDS methodology. In section 3, we explore VMAF's response to these distortions and propose a correction to improve VMAF's localized quality sensitivity and prediction behavior.

2. ESTIMATING THE PERCEIVED DISTORTIONS

The effect of distortion on the human perception can be estimated with MLDS[13, 14]. This methodology is effective in the process of selecting stimuli to compare in subjective study. Moreover, in subjective quality assessment of image/video, it has been shown that Two-Alternative Forced Choice (2AFC) methods are more precise and sensitive in comparison with direct rating methods. We applied MLDS methodology to estimate supra-threshold differences in small videos, *tubes*, encoded using AV1.

We want to estimate the perceptual distance in a *tube-content* C_i, a set of video-tubes composed of a reference tube and 5 levels of distortions: $C_i = \{S_1^i, S_2^i, ..., S_6^i\}$. The stimuli are pre-ordered in increasing value of quantization parameters (QP), with the assumption that higher QPs introduce higher perceptual distortions: $S_1^i < S_2^i < ... < S_6^i$.

In the original work of MLDS, a *perceptual curve* per content is estimated from the judgments collected. The limitation is that the perceptual curves of 2 contents cannot be directly related. To compare them and estimate a scale factor for each of the perceptual curves, we added inter-content comparisons. Where a quadruplet is composed of a pair of stimuli from content C_i and a pair from content C_j: $(S_a^i, S_b^i, S_c^j, S_d^j)$.

More details about the solving procedure and the selection of quadruplets for inter-content comparisons are in [15]. In the next sections, we will present how we constructed and selected our *tube-contents* to be presented in the subjective experiment.

[1]AV1 encoder v3.1.2, from AOM Alliance Open Media: https://aomedia.googlesource.com/aom/

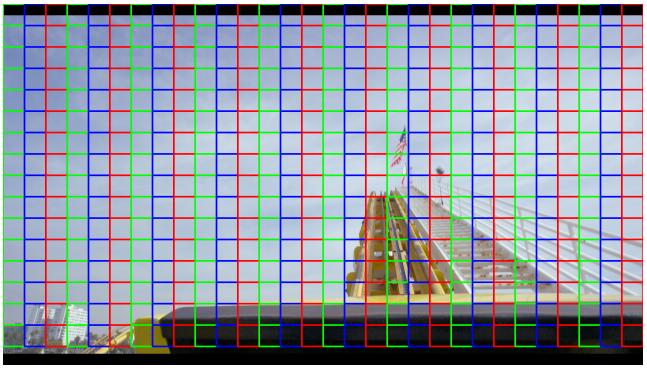

Fig. 1: Tubes extraction process on "videoSRC057" from VideoSet dataset [16], a frame is divided in 30×16 tubes.

2.1. Dataset creation

We used the reference videos from VideoSet database [16]. Particularly, the subset of 115 reference videos with a size of 1080p and at 30fps was selected. We encoded them using AV1 encoder. During the encoding, we varied the QP values, ranging from 3 to 63 with a step of 2, generating 31 Processed Video Sequences (PVS) for each original video sequence (SRC).

From the original video we extracted *tubes* fig. 1. Tubes design is based on human perception, with the length of the tube modeling an average gaze duration (400ms): the fixation performed by the eyes on an object. The spatial size of the tube is derived from the size of the human fovea: covering 1° at the center of the field of view, around 64 pixels in standard visualization conditions. Different Group of Pictures (GOPs) sizes can potentially influence the perceptual distortions, nevertheless, the selected parameters in the encoding algorithm did lead to a relatively homogenous distortion for the whole tube.

A *tube-content* in our case is a set of *tubes*: a first *tube* extracted in the reference video of the SRC, and 31 distorted versions of that *tube* from the 31 PVS.

2.2. Tube-contents clustering

To select interesting and diverse *tube-contents* to evaluate through subjective experiments, we choose to qualify them using Full-Reference quality metrics, e.g. VMAF[17], SSIM[4], VIF[18], DLM[19], PSNR and also the perceptual deep learning based metric LPIPS[20]. For each tube-content, we gather a list of 31 scores per metric, each score corresponding to each distortion level (PVS).

In the figure 2, the relation between 2 metrics is extracted using linear regression. The slope, intercept and RMSE of the fitting are then used to represent how objective metrics qualify a tube-content. Moreover, by coupling metrics together and extracting fitting parameters, we can get information about the difference in sensitivity and agreement of the objective quality metrics on the same tube-content.

The 6 quality metrics mentioned above, give 15 metric couples with 3 features each: slope, intercept, and RMSE of

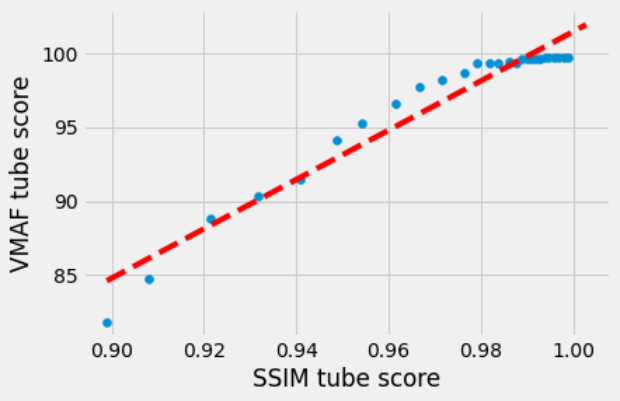

Fig. 2: Fitting a tube-content with VMAF/SSIM scores. 31 blue dots, one for each distortion level (e.g QP value), scattered based on SSIM and VMAF scores. In red the best linear fit line, with slope 166.45, intercept 65.008, and RMSE 5.906.

best fit line. The 45 resulting features are extracted for all the tube-contents constructed in the previous section and used to cluster them. We applied K-means clustering algorithms on the 45 features collected. Empirically, we found that 96 clusters were modelling the data correctly. In the figure 3 examples of tubes from 4 clusters are displayed. We can see that similar tube-contents are clustered together.

Fig. 3: Examples of tube-contents from 4 clusters, each tube-content reference tube is represented, noise pattern is used as right-padding to square visualizations. Top clusters tend to capture high spatial and temporal frequency texture, while bottom clusters model more flat or darker textures.

2.3. Tube-contents selection for subjective experiments

The tube-contents to annotate via subjective experiment are selected from clusters with specific properties: clusters having a large number of samples, and where VMAF and SSIM disagree (i.e, large RMSE, and small or negative slope). These specific contents can help us to better correct VMAF

Fig. 4: The 20 tube-contents selected for the subjective experiment. Each column represents a tube-content and its distortion levels in increasing order from the top with the original tubes (not distorted) to bottom with the most distorted level.

on the local quality level.

For our first study, we selected 20 tube-contents, each tube-content contains 32 tubes (1 reference + 31 distorted versions of it), and we need to reduce this number to have manageable number of quadruplets to annotate. We selected 6 distortion levels to conduct the subjective experiment via MLDS methodology. The final selection is depicted in fig. 4.

2.4. Tube-contents distortion level selection.

To select interesting distortion levels to evaluate during the subjective experiment, we use VMAF scores as an estimate of the distortion. For each tube-content, we first computed the range from VMAF scores between the QP0 and QP55 of AV1 encoded tubes and divided it by $N = 6$ in our case, yielding a *step score* uses to select, $N - 1$ equally spaced QPs on the VMAF scale, starting from the reference tube also included.

2.5. Quadruplet selection: intra and inter annotations

The quadruplets for intra-contents estimation via subjective experiment are generated following the recommendation of MLDS papers [13, 14], where there are $\binom{N}{4}$ quadruplets for a N levels difference scaling experiment. In our case we selected $N = 6$ levels to estimate (a reference tube + 5 levels of distortion) yielding 15 quadruplets to evaluate for each tube-content.

In the case of inter-contents comparisons, we can create $\frac{N\times(N+1)}{2}$ pairs for the content A to potentially be compared with the $\frac{N\times(N+1)}{2}$ pairs of a content B, which is not tractable with a growing number of content. Instead, we use the approach described in [15] to reduce the amount of comparisons and increase our efficiency in the collection of subjective data.

2.6. In lab subjective test results

25 expert observers were recruited to annotated first the intra-quadruplets of the 20 tube-contents selected using the method described in previous sections. The annotation procedure was divided into 10 sessions of 30 trials each, lasting on average 6 minutes. Each observer performed the annotation procedure over a 2-week period, resulting in 300 annotations per observer and 375 annotations per tube-content. In fig. 5, some examples of estimated intra-content perceptual curves from the subjective data.

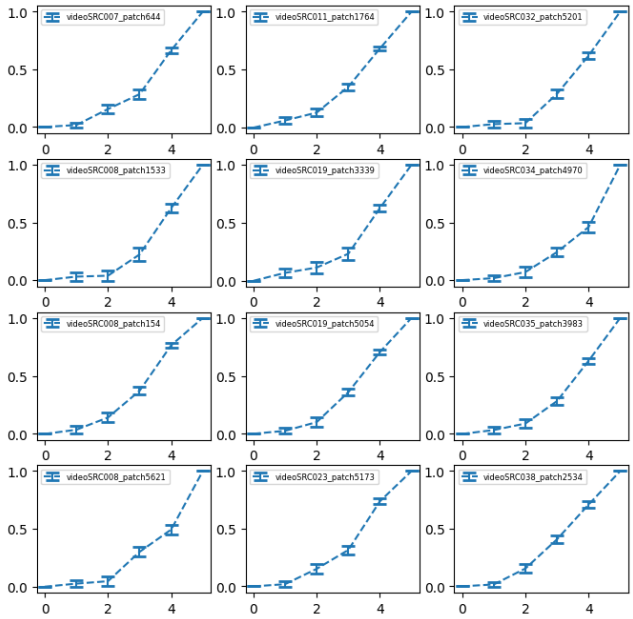

Fig. 5: Estimation of 12 individual perceptual curves from the intra-quadruplet dataset. Estimation performed with Maximum Likelihood-based solving and standard deviation obtained via bootstrapping over the data [21], on the x-axis the distortion levels indices of each tube-content, 0: reference tube, 5: the most distorted level, and the y-axis the estimated scores.

In addition, we collected inter-content difference scaling judgments, in total 7500, by sampling boosting strategy described in [15]. Estimation of the perceptual curves and their scaling inter-content can be seen in fig. 6. The contents of the tube "wall-spot" and "rain", in orange and gray, are judged to introduce a lot of distortion, from level 3 where they are above all other contents on the perceptual scale.

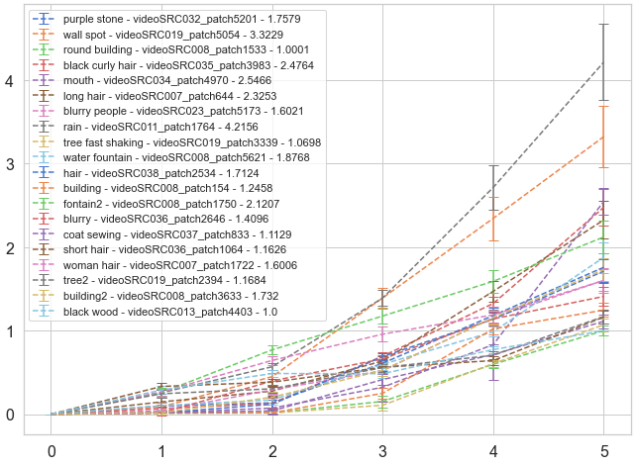

Fig. 6: Scaling estimation between the 20 perceptual curves with inter-content quadruplet annotations. Estimation performed with Maximum Likelihood-based solving and standard deviation obtained via bootstrapping over the data [21], on the x-axis the distortion levels indices like fig. 5, and the y-axis the estimated scores.

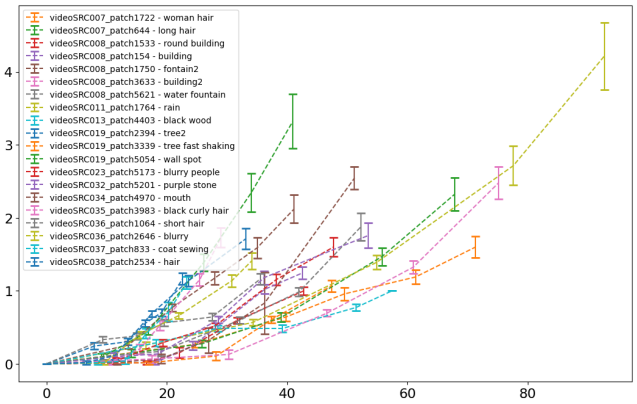

Fig. 7: Comparison of the 20 perceptual curves with VMAF. On the x-axis are the VMAF scores (i.e, 100 - VMAF), where near 0 scores indicate no perceived distortion, and the y-axis represents the estimated perception scores.

3. VMAF CORRECTION FOR LOCAL QUALITY

In this section, we present how we propose to correct VMAF for a better estimation of the local quality in video, using the subjective data collected in the previous section. We observe that VMAF tends to overestimate the visibility of low level distortions: an oversensitivity for small distortions, see fig. 7, where x-axis values, VMAF, increase faster than y-axis values, subjective scores. We made the same observation for other metrics: PSNR, SSIM, DLM, VIF and LPIPS.

Based on the subjective data retrieved, we can first evaluate the different objective metrics, table 1. We reported Pearson correlation coefficient (PLCC), Kendall tau correlation (KRCC), and Spearman correlation (SRCC). We can see that VMAF has the highest correlation with the estimate scores. Despite the good performance of VMAF, there is still room for improvement.

Moreover, we analyze the performance in different ranges of distortion, table 2. The low distortion are equivalent to low QP values and high fidelity encoded tubes. It is interesting to see again that VMAF is the most precise in low distortion levels but in high distortion levels LPIPS becomes better.

	PLCC	KRCC	SRCC
VMAF [17]	**0.8236**	**0.7292**	**0.8958**
DLM [19]	0.7953	0.6843	0.8614
VIF [18]	0.6705	0.6685	0.8458
SSIM [4]	0.7425	0.6691	0.8401
PSNR	0.5496	0.6406	0.8025
LPIPS (AlexNet) [20]	0.8074	0.7096	0.8817
LPIPS (SqueezeNet)	0.4920	0.6875	0.8555

Table 1: Performances of Full-Reference metrics on the dataset.

Using the subjective data, we proposed to retrain VMAF, the SVM pooling of metrics, training on the data from 12 tube-contents (i.e: 72 data points), and testing on the remaining 8 (i.e: 48 data points). To report performances of the retraining, we averaged the result over a set of 1000 permutations of train/test set splitting over the 20 tube-contents, in table 3.

range		PLCC	KRCC	SRCC
0 to 0.6	VMAF [17]	**0.6303**	**0.4645**	**0.6414**
	SSIM [4]	0.4523	0.3388	0.5063
	PSNR	0.4078	0.3159	0.4547
	LPIPS (AlexNet) [20]	0.5995	0.4204	0.5802
0.6 to +inf	VMAF [17]	0.6490	0.3985	0.5629
	SSIM [4]	0.5428	0.2882	0.4010
	PSNR	0.3724	0.2996	0.4104
	LPIPS (AlexNet) [20]	**0.7215**	**0.4351**	**0.6165**

Table 2: Performance of Full-Reference metrics on the different distortion ranges. The dataset is split into 2 equal sized subsets.

	PLCC	KRCC	SRCC
VMAF [17]	0.8685	0.7433	0.9015

Table 3: Performances after retraining of VMAF: average score over 1000 permutations of 8 tube-contents in the test set.

4. CONCLUSION AND FUTURE WORK

In this work, we studied the behavior of different quality metric at a localized scale in videos. We showed that VMAF is a good candidate to estimate these distortions. Using the subjective data collected with MLDS methodology, we proposed a correction of VMAF to estimate the local quality in video content encoded using AV1. We plan to extend this dataset to more tubes in a large-scale crowdsourcing study to increase robustness to different contents. With a larger amount of data available it will be interesting to see if we can correct even more and maybe improve the pooling strategy of VMAF on the local spatial and temporal horizon. Another important aspect will be to also use this metric in CODEC to see if it can improve the encoding of videos.

5. REFERENCES

[1] Christophe Charrier, Laurence T Maloney, Hocine Cherifi, and Kenneth Knoblauch, "Maximum likelihood difference scaling of image quality in compression-degraded images," *JOSA A*, vol. 24, no. 11, pp. 3418–3426, 2007.

[2] Christophe Charrier, Kenneth Knoblauch, Laurence T Maloney, and Alan C Bovik, "Calibrating ms-ssim for compression distortions using mlds," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 3317–3320.

[3] Christophe Charrier, Kenneth Knoblauch, Laurence T Maloney, Alan C Bovik, and Anush K Moorthy, "Optimizing multiscale ssim for compression via mlds," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4682–4694, 2012.

[4] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[5] Vlado Menkovski, Georgios Exarchakos, and Antonio Liotta, "Adaptive testing for video quality assessment," *Quality of Experience of multimedia content sharing, Lisbon, Portugal*, 2011.

[6] Vlado Menkovski, Georgios Exarchakos, and Antonio Liotta, "Tackling the sheer scale of subjective qoe," in *International Conference on Mobile Multimedia Communications*. Springer, 2011, pp. 1–15.

[7] Vlado Menkovski, Georgios Exarchakos, and Antonio Liotta, "The value of relative quality in video delivery," *Journal of Mobile Multimedia*, pp. 151–162, 2011.

[8] Vlado Menkovski and Antonio Liotta, "Adaptive psychometric scaling for video quality assessment," *Signal Processing: Image Communication*, vol. 27, no. 8, pp. 788–799, 2012.

[9] Antonio Liotta, Decebal Constantin Mocanu, Vlado Menkovski, Luciana Cagnetta, and Georgios Exarchakos, "Instantaneous video quality assessment for lightweight devices," in *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*, 2013, pp. 525–531.

[10] Karam Naser, Vincent Ricordel, and Patrick Le Callet, "Modeling the perceptual distortion of dynamic textures and its application in hevc," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3787–3791.

[11] Karam Naser, Vincent Ricordel, and Patrick Le Callet, "A foveated short term distortion model for perceptually optimized dynamic textures compression in hevc," in *2016 Picture Coding Symposium (PCS)*. IEEE, 2016, pp. 1–5.

[12] Hui Men, Hanhe Lin, Mohsen Jenadeleh, and Dietmar Saupe, "Subjective image quality assessment with boosted triplet comparisons," *IEEE Access*, vol. 9, pp. 138939–138975, 2021.

[13] Laurence T Maloney and Joong Nam Yang, "Maximum likelihood difference scaling," *Journal of Vision*, vol. 3, no. 8, pp. 5–5, 2003.

[14] Kenneth Knoblauch, Laurence T Maloney, et al., "Mlds: Maximum likelihood difference scaling in r," *Journal of Statistical Software*, vol. 25, no. 2, pp. 1–26, 2008.

[15] Andréas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, and Patrick Le Callet, "Improving maximum likelihood difference scaling method to measure inter content scale," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2045–2049.

[16] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeong-Hoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, Yun Zhang, Jiwu Huang, Sam Kwong, and C.-C. Jay Kuo, "Videoset: A large-scale compressed video quality dataset based on JND measurement," *CoRR*, vol. abs/1701.01500, 2017.

[17] Netflix, "Vmaf v0.6.1 model," https://github.com/Netflix/vmaf.

[18] Hamid R Sheikh and Alan C Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.

[19] Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.

[20] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *CoRR*, vol. abs/1801.03924, 2018.

[21] Felix A Wichmann and N Jeremy Hill, "The psychometric function: Ii. bootstrap-based confidence intervals and sampling," *Perception & psychophysics*, vol. 63, no. 8, pp. 1314–1329, 2001.