# VMAF Reproducibility: Validating a Perceptual Practical Video Quality Metric

Reza Rassool
CTO
*RealNetworks*
Seattle, WA, US
reza@realnetworks.com

*Abstract*— **Measuring video quality with standard metrics ensures that operators can deliver to consumers the desired quality of experience (QoE) at an optimal cost. Such metrics also allow CODEC engineers to optimize the performance of their encoding algorithms. This paper briefly surveys existing video quality metrics and then presents results of the new Video Multi-Method Assessment Fusion (VMAF) metric [1] proposed by Netflix. The author and colleagues used VMAF to measure the quality of a 4K dataset encoded with the RealMedia video CODEC at a range of bitrates. They also gathered subjective quality assessments from a group of viewers for the same dataset. The paper presents findings of correlation between subjective and objective results.**

## I. Introduction

In recent years, researcher Staelens et al [2], stated that while “subjective quality assessment” measured with mean opinion score (MOS) “is the most accurate method for obtaining human judgments on video quality,” they lament that subjective evaluations “are time-consuming, expensive, and require specialized expertise.” Also, “they cannot produce real-time quality ratings throughout a distribution network.” This echoes the findings of several researchers evaluating objective QoE metrics and the industry has been in search of an objective video quality metric that is easy to compute and that matches the expensive ‘goldeneyes’ subjective tests.

## II. Survey of various objective metrics:

### A. PSNR (Peak Signal to Noise Ratio)

PSNR is a traditional signal quality metric, measured in decibels. It is directly derived from mean square error (MSE) or its square root (RMSE). The formula used is:

$$20 * \log10 ( Max / RMSE ) \qquad (1)$$

where the error is computed over all the pixels in the video with respect to a reference video. PSNR successfully provides a numerical value when comparing an original input file with a coded output file, but doesn’t match the ‘golden eyes’ tests when humans are put in a room.

Variants of PSNR include:

*Frame-averaged PSNR* includes a temporal component to the content comparison.

*PSNR-HVS-M* [3] applies the comparison in the frequency domain. Good correlation with MOS results are claimed by researcher Daede et al with appropriate selection of weighting factors to the DCT coefficients.

### B. SSIM (Structural Similarity Image Metric)

SSIM is a still image quality metric introduced in 2004 by researcher Wang et al [4]. 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 with respect to a reference image. The original paper produces scores ranging between zero and one; however, this is commonly expressed in a non-linear decibel scale:

$$-10 * \log10 (1 - SSIM) \qquad (2)$$

Algorithmic variant, MS-SSIM, handles multi-scale windows while FMS-SSIM is a fast implementation with constrained scales. This metric has been adapted for streamed video and commercialized as SSIMWave.

### C. VQM (Video Quality Monitor)

VQM is a commercial tool that requires no reference but measures the video quality with respect to the format and protocol specification. It can objectively measure blockiness, blurriness and frame rate.

### D. VMAF (Video Multi-Method Assessment Fusion)

VMAF was specifically formulated by Netflix to correlate strongly with subjective MOS scores. Using machine learning techniques, a large sample of MOS scores were used as ground truth to train a quality estimation model. It is a full-reference, perceptual video quality metric that aims to approximate human perception of video quality. This metric is focused on quality degradation due to compression and rescaling. VMAF estimates the perceived quality score by computing scores from multiple quality assessment algorithms and fusing them using a support vector machine (SVM). Currently, three image fidelity metrics and one temporal signal have been chosen as features to the SVM:

*a)* Anti-noise SNR (AN-SNR),
*b)* Detail Loss Measure (DLM),
*c)* Visual Information Fidelity (VIF)

#### *d)* Mean Co-Located Pixel Difference (MCPD)

The MCPD of a frame with respect to the previous frame, the temporal component, is important and is often lacking in metrics that merely compare a decoded image with a reference image. The rate control mechanism of a CODEC is continuously making a trade-off between spatial and temporal quality. VMAF is a good metric by which that compromise can be judged.

VMAF does not consider color separately.

## III. Reproducing VMAF with RealMedia CODEC at 4K

### A. Method

4K video clips were encoded at a range of bitrates. Subjects scored each encoded clip for quality with respect to the reference original. The VMAF score for each clip was then computed.

1) Content

Ten video clips (listed in the legend of fig. 1) from a Xiph dataset [VQEG4K] were encoded at bitrates from 3mb/s to 10mb/s using the RealMedia CODEC, rmXD.

2) MOS

Eighteen subjects provided opinion scores (as defined in *table 1.)* using double stimulus standardized testing as defined in REC-BT.500 [5]. Normalized Differential Mean Opinion Scores (DMOS) were computed as described in [1].

3) VMAF

Each reference file was encoded from YUV then decoded to YUV and its VMAF score was computed.

### B. Results

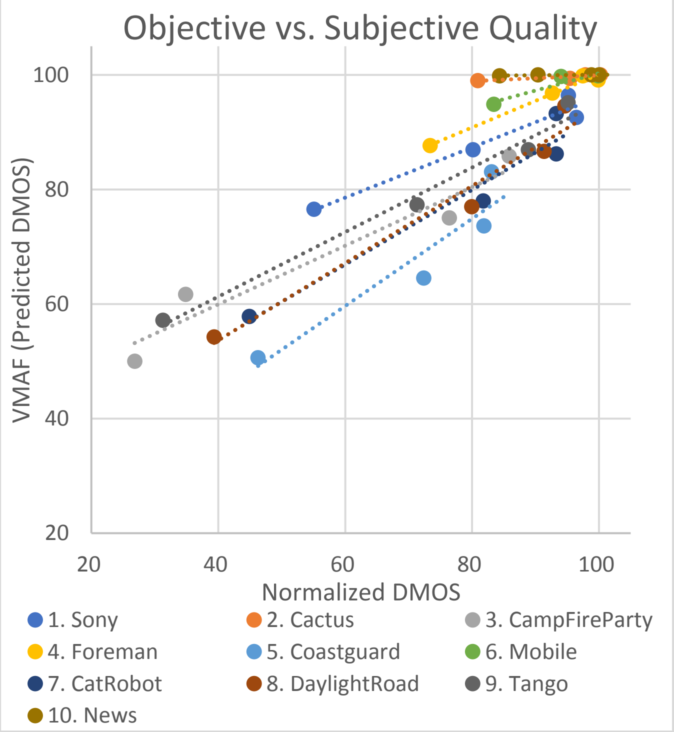

*Fig. 1 Chart of VMAF vs. DMOS for 4K video dataset*

| Score | Opinion of quality compare to original |
|---|---|
| 1 | Very Annoying |
| 2 | Annoying |
| 3 | Slightly annoying |
| 4 | Noticeable but not annoying |
| 5 | Indistinguishable from original |

*Table 1. Opinion Scores*

## IV. Conclusion

*a)* The results exhibit a strong correlation between subjective Mean Opinion Score and the computed objective VMAF score with a correlation of 0.948. This compares well with the Netflix results of 0.963 and 0.939 for the NFLX-TEST and VQEGHD3 datasets, respectively. Even though the original VMAF was trained on 1080p data we demonstrate its applicability to 4K video.

*b)* While VMAF is a strong predictor of subjective opinion of a collection of viewers on the quality of video content, with a RMSE of 12.7, our results show that the computed value more often (85 percent of the time) over-estimates the subjective quality.

*c)* The results indicate that if a video service operator were to encode video to achieve a VMAF score of about 93 then they would be confident of optimally serving the vast majority of their audience with content that is either *indistinguishable from original* or with *noticeable but not annoying* distortion.

## Acknowledgment

Thanks to colleagues Xiaobo Liu, *Senior Staff* and Chia-Yang Tsai, *Chief Architect,* and Qiang (Joe) Li, *Director of Engineering* and volunteer testers.

## References

[1] Zhi Li, Anne Aaron et al, “Toward A Practical Perceptual Video Quality Metric,” *Netflix TechBlog, June, 2016*

[2] Staelens, Nicolas, et al. "Measuring video quality in the network: from quality of service to user experience." *9th International Workshop on Video Processing and Consumer Electronics (VPQM 2015)*. 2015.

[3] Nikolay Ponomarenko et al, “On between-coefficient contrast masking of DCT basis functions,” *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-07, January, 2007, 4p.*

[4] Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity." *IEEE transactions on image processing* 13.4 (2004): 600-612.

[5] ITU, “Methodology for the subjective assessment of the quality of television pictures,” Recommendation ITU-R BT.500-13.

[6] Elemental, http://www.elemental.com/resources/4k-test-sequences, MediaLab, Shanghai Jiao Tung University, http://medialab.sjtu.edu.cn/web4/index.htm