A machine learning driven solution to the problem of perceptual video quality metrics

Stamos Katsigiannis, Hassan Rabah, Naeem Ramzan

Chapter 12

A machine learning driven solution to the problem of perceptual video quality metrics

Stamos Katsigiannis[1], Hassan Rabah[2] and Naeem Ramzan[1]

The advent of high-speed internet connections, advanced video coding algorithms, and consumer-grade computers with high computational capabilities has led video-streaming-over-the-internet to make up the majority of network traffic. This effect has led to a continuously expanding video streaming industry that seeks to offer enhanced quality-of-experience (QoE) to its users at the lowest cost possible. Video streaming services are now able to adapt to the hardware and network restrictions that each user faces and thus provide the best experience possible under those restrictions. The most common way to adapt to network bandwidth restrictions is to offer a video stream at the highest possible visual quality, for the maximum achievable bitrate under the network connection in use. This is achieved by storing various pre-encoded versions of the video content with different bitrate and visual quality settings. Visual quality is measured by means of objective quality metrics, such as the Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Visual Information Fidelity (VIF), and others, which can be easily computed analytically. Nevertheless, it is widely accepted that although these metrics provide an accurate estimate of the statistical quality degradation, they do not reflect the viewer's perception of visual quality accurately. As a result, the acquisition of user ratings in the form of Mean Opinion Scores (MOS) remains the most accurate depiction of human-perceived video quality, albeit very costly and time consuming, and thus cannot be practically employed by video streaming providers that have hundreds or thousands of videos in their catalogues. A recent very promising approach for addressing this limitation is the use of machine learning techniques in order to train models that represent human video quality perception more accurately. To this end, regression techniques are used in order to map objective quality metrics to human video quality ratings, acquired for a large number of diverse video sequences. Results have been very promising, with approaches like the Video Multimethod Assessment Fusion (VMAF) metric achieving higher correlations to user-acquired MOS ratings compared to traditional widely used objective quality metrics.

[1]University of the West of Scotland, UK
[2]University of Lorraine, France

In this chapter, we examine the performance of VMAF and its potential as a replacement for common objective video quality metrics.

12.1 Introduction

The last decade saw an explosion in the use of video streaming platforms, with predictions estimating that video streaming will account to 82% of global IP traffic by 2022, up from 75% in 2017 [1]. Available network capacity keeps increasing both for the traditional wired land-based networks and for the wireless mobile networks. Optic fibre-based networks can offer bandwidths in the range of Gbps, while next generation 5G mobile networks promise speeds up to 10 Gbps and latency of less than 1 ms. Despite the phenomenal increase in network capacity, demand for bandwidth continues to increase as video content of higher resolutions becomes available. Streaming raw video is unpractical due to extremely high bandwidth requirements. For example, streaming 1 second of a 1920×1080 video sequence at 30 fps with 24 bit colour depth would require the transmission of approximately 178 MB of data, without including protocol overheads. Furthermore, network capacity and latency have large variations depending on the user, the location, the time within the day, and overall network congestion, while the capability of end-user devices to play the transmitted videos depends on their hardware specifications. Sophisticated video compression and transmission algorithms are used in order to address these issues and facilitate the streaming of high quality video in practice.

The most common approach for achieving flexible and uninterrupted video streaming under variable network conditions is to offer to the users the highest quality video stream possible under the specific conditions and constraints of each user. To achieve this, video sequences are compressed using various bit rates and/or spatial resolutions and each version is divided in smaller segments that are typically a few seconds long. Then, these video segments are made available via a web server and are accessed by video player software using standard HyperText Transfer Protocol (HTTP) GET requests. A manifest file containing information about all the available versions of the video is created and used by the client in order to request the segments that fit its requirements. Requirements depend on bandwidth, latency, hardware capabilities, etc. and can vary during playback. As a result, adaptation to different versions is achieved by requesting different segments of a video, thus increasing or decreasing the received quality as seen fit. While various video streaming protocols exist, the most dominant are Apple HTTP Live Streaming (HLS) [2] and MPEG Dynamic Adaptive Streaming over HTTP (DASH) [3].

Although video compression allows the transmission or storage of video content in a practical and cost-effective manner, it suffers from an important side effect. The reduction in bit rate or in spatial resolution used for the compressed sequences leads to a reduction in visual quality. The compressed video sequences can suffer from loss of fidelity, compression artefacts, colour fading and blur among others. Considering that video streaming systems aim to stream to the users the highest quality video possible, the accurate evaluation of the visual quality of each compressed version of the video becomes crucial, especially the perceptual quality as experienced by the

human visual system (HVS) [4]. Video Quality Assessment (VQA) techniques can be divided into two categories, depending on how the quality evaluation is achieved: (a) Objective VQA approaches, and (b) Subjective VQA approaches.

Objective video quality models consist of mathematical models that measure the differences between the original and the distorted video sequences and deterministically output the same quality score when the same video sequences are compared under the same parameters. Objective VQA approaches are further divided into three categories, depending on the amount of information they require regarding the original video sequence:

1. **Full Reference (FR)**. FR approaches require full access to the reference video sequence in order to estimate the quality of the examined version.
2. **Reduced Reference (RR)**. RR approaches require only a set of features extracted from the reference video sequence in order to estimate the quality of the examined version.
3. **No Reference (NR)**. NR approaches require no information regarding the reference video sequence.

Subjective VQA approaches measure video quality as perceived by humans. Video quality scores are computed by conducting experiments where users are asked to watch and rate the quality of various video sequences. The acquired ratings are then averaged in order to compute the final video quality rating, as perceived by the human viewers, in the form of a Mean Opinion Score (MOS). Subjective VQA approaches can be divided in two categories, depending on how the quality rating experiments are conducted:

1. **Single-Stimulus**. In Single-Stimulus experiments, viewers are requested to rate the quality of single video sequences, without having access to the reference video sequences.
2. **Double/Multiple-Stimulus**. In Double/Multiple-Stimulus experiments, viewers watch both the reference and distorted versions of the video sequences and provide ratings in relation to the reference.

It is well accepted in the literature that objective quality metrics fail to accurately model video quality as perceived by human viewers. To this end, subjective video quality ratings provide the most accurate measurement of video quality. However, conducting subjective video quality rating experiments to acquire MOS ratings is a time-consuming, expensive, and arduous task, thus it cannot be practically employed by video streaming providers that have hundreds or thousands of videos in their catalogues. Researchers have tried to bridge the gap between objective and subjective VQA by training machine learning models in order to map objective quality metrics to subjective quality ratings (MOS), effectively using the easy to compute deterministic objective metrics to predict ratings subjectively decided by human viewers. This chapter examines the performance of one of the most promising machine learning-based VQA metrics, the Video Multimethod Assessment Fusion (VMAF) [5].

12.2 Objective Video Quality Assessment Methods

Various FR, RR, and NR VQA methods that do not rely on machine learning have been proposed in the literature across the years, with some, like the PSNR and the SSIM, being extensively used by the industry and the research community in production systems and as benchmarks. An overview of some of these traditional metrics is provided in this section.

12.2.1 Full-Reference metrics

Some of the most commonly used FR VQA methods were originally used for the quality assessment of images and were extended for use in video by being applied to each frame of the videos. Methods like that include the Mean Squared Error (MSE)-based Peak Signal-to-Noise Ratio (PSNR), the Visual Information Fidelity (VIF) [6], the Structural Similarity (SSIM) index in its various forms, [7] and the Visual Signal-to-Noise Ratio (VSNR) [8]. However, it is well established that such metrics fail sometimes in characterising the perceptual quality of the video sequences depending on the types of the distortion and the content of the video [9, 10, 11]. Various other FR metrics have been proposed in the literature [12]. Aabed et al. [13] proposed a perceptual quality metric that utilises low complexity power spectral features in the frequency domain, while Manasa and Channappayya [14] proposed the use of optical flow statistics, such as the minimum eigenvalue of the local flow patches, the mean, the standard deviation, and the coefficient of variation in order to estimate temporal quality and the use of SSIM for spatial quality estimation, combining both for computing the final quality score. Seshadrinathan and Bovik [15] proposed the MOVIE index that examines temporal, spatial, and spatiotemporal characteristics of distortion in order to estimate video quality. Various works also examined video quality in relation to motion [15], in relation to the frame rate and quantisation [16], and in relation to network QoS and application QoS within the context of web streaming [17].

12.2.2 Reduced-Reference metrics

The biggest difficulty in designing RR metrics is the extraction of suitable and descriptive features from the reference video sequence in order to have sufficient data for an accurate video quality prediction [18]. Tao et al. [19] proposed a relative video quality metric, rPSNR, that can be computed without parsing or decoding the transmitted video, and without any knowledge of video characteristics. Piamrat et al. [20] proposed the Pseudo Subjective Quality Assessment (PSQA) metric, a hybrid metric that makes use of objective and subjective features to evaluate QoE for video streaming in wireless networks. Entropic differences and wavelet-based natural video statistics were utilised by Soundararajan and Bovik [21] for their RR video quality metric, while Baik et al. [22] used a machine learning model to estimate the effect of spatial distortions, types of buffering and resolution changes in quality degradation.

12.2.3 No-Reference metrics

Various NR metrics have been proposed in the literature following different approaches [9, 23]. The NORM algorithm by Naccari et al. [24] uses macroblock information at the decoder level in order to evaluate distortions on H.264/AVC streamed videos. Wu et al. [25] used local texture and global intensity features extracted from the decoded video to evaluate the quality of stalled streaming video. Pixel-based features and bitstream information was used by Winkler and Mohandas [12] for real-time video quality estimation, while a combination of a spatio-temporal natural scene statistics (NSS) model for videos and a motion model that quantifies motion coherency in video scenes was proposed by Saad et al. [26]. Mittal et al. [27] exploited the intrinsic statistical regularities observed in natural videos to achieve NR quality assessment, while features based on a 3D shearlet transform were used by Li et al. [28].

12.3 The Video Multimethod Assessment Fusion (VMAF) Metric

Expanding on previous work by Liu et al. [29] and Lin et al. [30], researchers developed the Video Multimethod Assessment Fusion (VMAF) [5] FR VQA metric that employs a machine learning approach in order to map multiple elementary objective quality metrics to subjective quality ratings (MOS). The rationale behind this approach is that although each individual objective metric cannot fully capture the perceptual quality of the video, as it has its respective drawbacks and advantages, "fusing" multiple metrics together by assigning weights to each through a machine learning algorithm could potentially preserve the advantages of each metric and deliver a more accurate final video quality score. Furthermore, since these weights would be trained for optimising the accuracy of predicting subjective ratings provided by actual viewers, it is expected that such a metric would be more accurate in predicting perceptual video quality. The VMAF metric was originally trained on the NFLX Video Dataset [5], while a newer subjective dataset with a broadened scope was used for recent releases [31] that included more diverse content and source artefacts such as film grain and camera noise, and a larger range of encoding resolutions and compression parameters.

VMAF uses Support Vector Machine (SVM) regression [32] to map the combination of the following three elementary metrics to subjective video quality ratings [5]:

- **Visual Information Fidelity (VIF)** [6], which is a widely used image quality metric. The original VIF metric measures quality by determining the loss of fidelity in four scales. VMAF used a modified version of VIF where the loss of fidelity at each of the four scales is considered as an elementary metric instead of the combined VIF score.
- **Detail Loss Measure (DLM)** [33], which is an image quality metric that measures the loss of details that affect content visibility. Although originally pro-

posed to be used in combination with the Additive Impairment Measure (AIM), VMAF uses only DLM as an elementary metric.
- **Motion** [5], which is a simple temporal feature that measures the average low-pass filtered differences between consecutive frames.

VMAF scores are computed for each frame of a video sequence and the final VMAF score for a video sequence is computed through simple temporal pooling by computing the arithmetic mean of the VMAF scores across all frames, as experiments showed that the arithmetic mean yields the highest correlation with subjective scores.

Experiments on various subjective video quality datasets showed that VMAF scores were better correlated to MOS ratings compared to traditional quality metrics [5] and the VMAF metric has now been adopted widely by industry and researchers [31]. VMAF scores have a range from 0 to 100, with 0 being the lowest and 100 the highest quality. The current default VMAF model (v0.6.1) has been trained using subjective ratings acquired using a 1080p display with a viewing distance of 3H, H being the height of the screen, and following an Absolute Category Rating (ACR) methodology were viewers rated the quality of the video sequences on the scale of "bad", "poor", "fair", "good", and "excellent". Under the specific viewing conditions that the subjective quality tests were conducted, a VMAF score of 20 maps the "bad" quality, a score of 100 the "excellent" quality, and a score of 70 would be between "good" and "fair" [31]. As a result, when applying the default VMAF model to video sequences of different spatial resolution than 1080p, then the VMAF scores refer to ratings at different viewing conditions. Considering that the default model measures quality at the critical angular frequency of $1/60$ degree/pixel, this geometry applies to 1080p at 3H, to 720p at 4.5H, to 480p at 6.75H, etc. As a result, the application of the default model to a 480p video sequence would yield a quality rating referring to a viewing distance of 6.75H [31]. When computing VMAF on down-sampled video sequences, VMAF developers suggest the up-sampling of the sequence to the resolution of the reference sequence before VMAF application as otherwise the obtained VMAF score will fail to capture scaling artefacts [31].

12.4 Experimental evaluation

To evaluate the performance of VMAF under various settings, VMAF (v0.6.1), Y-PSNR, SSIM, and MS-SSIM quality scores were computed for the video sequences of three publicly available video datasets and their Pearson's Correlation Coefficient (ρ) with the available MOS ratings, as well as the R^2 for the linear fit were computed.

12.4.1 Datasets

12.4.1.1 MPEG-JVET2018 video sequences dataset

The MPEG-JVET2018 test video sequences contained five test sequences at a resolution of 1080p (1920×1080), progressively scanned using 4:2:0 colour sampling, with a duration of 10 seconds. Two sequences had a frame rate of 50 fps and 8 bits per sample, one sequence 60 fps and 8 bits per sample, and two sequences 60 fps and 10 bits per sample. Details of the test sequences are provided in Table 12.1.

Table 12.1 Details of the MPEG-JVET2018 dataset's video sequences. The resolution for all sequences is 1920×1080.

Name	Frame rate	Bit depth	Source	Target bit rate (kbit/s)			
				Rate 1	Rate 2	Rate 3	Rate 4
BQTerrace	60	8	NTT DOCOMO Inc.	400	600	1000	1500
RitualDance	60	10	Netflix	900	1500	2300	3800
MarketPlace	60	10	Netflix	500	800	1200	2000
BasketballDrive	50	8	NTT DOCOMO Inc.	800	1200	2000	3500
Cactus	50	8	EBU/RAI	500	800	1200	2000

The video sequences were encoded using two software packages, the HM 16.16 and the Joint Exploration Test Model (JEM) 7.0 software package [34]. The Joint Video Exploration Team (JVET) maintains the JEM software package in order to study coding tools in a coordinated test model [35]. The purpose of this datasets was to facilitate testing in accordance with BT.500 [36] in order to examine the proposals from 24 proponents. As a result, the MPEG-JVET2018 dataset contained (4 rates x 2 encoders x 24 proponents) x 5 sequences + 5 reference sequences = 965 video sequences in total. The MOS ratings were acquired using a degradation category rating (DCR) [37] method, leading to a quality rating scale with values in the range between 0 (lowest quality) and 10 (highest quality). Since all the video sequences in the MPEG-JVET2018 dataset are 1080p, they adhere to the specifications of the default VMAF model.

12.4.1.2 HEVC verification dataset (Tan et al. [38])

Twenty video sequences that have been used in [38] for the evaluation of video quality and compression performance of the H.265/HEVC [39] standard were used for the evaluation of the VMAF metric. Four categories of spatial resolutions were included in the experimental evaluation, namely UHD (3840×2160 except for one sequence that was 4096×2048), 1080p (1920×1080), 720p (1280×720), and 480p (832×480), with five video sequences per resolution. These sequences were selected from different sources so as to have different spatio-temporal characteristics, leading to differences in the behaviour of the compression algorithms utilised. Furthermore, the frame rate of the sequences spans from 30 to 60 frames per second, while all the sequences are in the $Y'C_BC_R$ colour space (as defined by ITU-R Rec. BT.709 [40]), with 8 bits per sample of each component. The video sequences were compressed using the AVC (JM-18.5, High profile [41]) and HEVC (HM-12.1, Main profile [42]) compression standards. For each sequence and each compression standard, four different fixed quantisation parameters (QP) settings were selected for compression so that the resulting bit rates for the respective HEVC-encoded sequences would be approximately half the bit rate of the AVC-encoded sequences, as well as so that their subjective quality would span a wide range of MOS values. Following this procedure, eight test sequences were created from each of the fifteen initial sequences resulting to a total of 160 test sequences. Furthermore, the quality of the created

video sequences in terms of average MOS was subjectively evaluated in [38] at two test sites, under a controlled laboratory environment. The MOS scores were recorded using a degradation category rating (DCR) [37] method, leading to a quality rating scale with values in the range between 0 (lowest quality) and 10 (highest quality). Only the 1080p sequences of the Tan et al. dataset adhere to the specifications of the default VMAF model. Nevertheless, it is interesting to examine VMAF's default model's performance against other metrics that are not constrained by resolution for video sequences of various resolutions.

12.4.1.3 ITS4S dataset

The ITS4S dataset [43] was primarily designed for the evaluation of NR video quality metrics and adheres to the following two factors: (a) the metric performance must degrade gracefully in response to new content (i.e. subject matter, camera, editing), and (b) the metric must accurately predict the quality of original videos (e.g., broadcast quality, contribution quality, professional cameras, prosumer cameras). It contains 813 video sequences from which 35% contain no compression artefacts, while the rest 65% contain simple impairments, in order to minimise the confounding factor of coding impairments on the original videos quality as the coding bitrate is reduced. The video content was selected out of a pool of HDTV and 4K videos [44] that were recorded using various resolutions and frame rates. The video sequences included in the ITS4S have been converted to 720p (1280×720) at 24 fps and were coded using H.264 High Profile VBR 2-pass coding at bitrates spanning from 0.512 Mbps to 2.340 Mbps, while 20 Mbps were used for the reference sequences. MOS ratings were acquired using the absolute category rating (ACR) method, leading to a quality rating scale with values in the range between 1 (lowest quality) and 5 (highest quality). Since VMAF is designed with coding and down-sampling related distortions in mind, its performance on the ITS4S dataset is expected to suffer since ITS4S contains numerous sequences with impairments that are unrelated to coding. Furthermore, although the resolution of the video sequences is not 1080p, both the reference and the impaired sequences have the same resolution (720p) thus no up-scaling is required. Nevertheless, as explained in section 12.3, the acquired VMAF scores will refer to a viewing condition of 4.5H distance from the screen.

12.4.2 Video quality scores

It must be noted that although MOS ratings were available for the reference sequences of the three datasets, the reference sequences were not included in the experimental comparison of the VMAF (v0.6.1), Y-PSNR, SSIM, and MS-SSIM metrics, since the Y-PSNR value for reference sequences is infinite, thus the correlation could not be established properly.

The performance of VMAF (v0.6.1), Y-PSNR, SSIM, and MS-SSIM in terms of the Pearson's Correlation Coefficient (ρ) and the R^2 in relation to the available MOS ratings for each dataset are provided in Table 12.2. Furthermore, Figures 12.1, 12.2, and 12.3 provide scatter plots showing the observers' MOS on the x-axis and the predicted score from the examined quality metrics on the y-axis. From Table 12.2,

Table 12.2 Pearson's Correlation Coefficient (ρ) and R^2 between the VMAF (v0.6.1), Y-PSNR, SSIM, and MS-SSIM scores and the available MOS for each dataset.

Dataset	VMAF		Y-PSNR		SSIM		MS-SSIM	
	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2
MPEG-JVET2018	**0.90**	0.816	0.70	0.492	0.89	0.792	**0.90**	0.807
Tan et al.	**0.66**	0.435	0.48	0.234	0.42	0.179	0.57	0.327
Tan et al. (1080p)	**0.72**	0.519	0.67	0.445	0.68	0.457	0.71	0.509
ITS4S	**0.34**	0.112	0.31	0.099	0.26	0.067	0.26	0.066

it is evident that VMAF achieved the best correlation with viewers' quality ratings for all the examined datasets. The performance of MS-SSIM was similar to VMAF for the MPEG-JVET2018 dataset ($\rho = 0.90$) and marginally worse for the 1080p sequences of the Tan et al. dataset ($\rho_{VMAF} = 0.72$ vs $\rho_{MS-SSIM} = 0.71$).

Regarding the MPEG-JVET2018 dataset, the correlation between VMAF scores and MOS was the highest among the three examined datasets (0.90). The video sequences in the dataset fully complied with the VMAF guidelines and model used (1080p with video coding-related distortions) and the accompanying MOS were relative to the reference video sequences. As expected, VMAF performed very well. Regarding the Tan et al. dataset, the correlation between VMAF and MOS was significantly lower (0.66) than the one achieved for the MPEG-JVET2018 dataset. Although the distortions in the video sequences were related to H264/AVC and H265/HEVC coding and the accompanying MOS were relative to the reference video sequences, the resolutions of the video sequences varied (UHD, 1080p, 720p, 480p). VMAF guidelines state that the default VMAF model is trained and optimised for 1080p video. When only the 1080p sequences of the Tan et al. dataset were examined, VMAFs correlation to MOS improved, albeit only slightly (0.72), with MS-SSIM achieving a marginally worse correlation of 0.71. Regarding the ITS4S datset, as expected due to the type of distortions included, the correlation between VMAF and MOS was the lowest among the examined datasets at 0.34, slightly better than the correlation with Y-PSNR which was 0.31. It seems that when MOS is not relative to the reference sequence and the scores do not scale similarly across different sequences, VMAF is not working well and Y-PSNR provides comparable performance as a quality metric. Interestingly, SSIM and MS-SSIM performed the worst for this dataset, achieving a correlation of 0.26. A larger and even more diverse dataset could help establish this argument more definitely, but still the evidence points towards that direction.

12.5 Conclusion

Examining the quality ratings achieved by VMAF, it is evident that VMAF scores are more aligned to viewer quality ratings compared to Y-PSNR, SSIM, and MS-

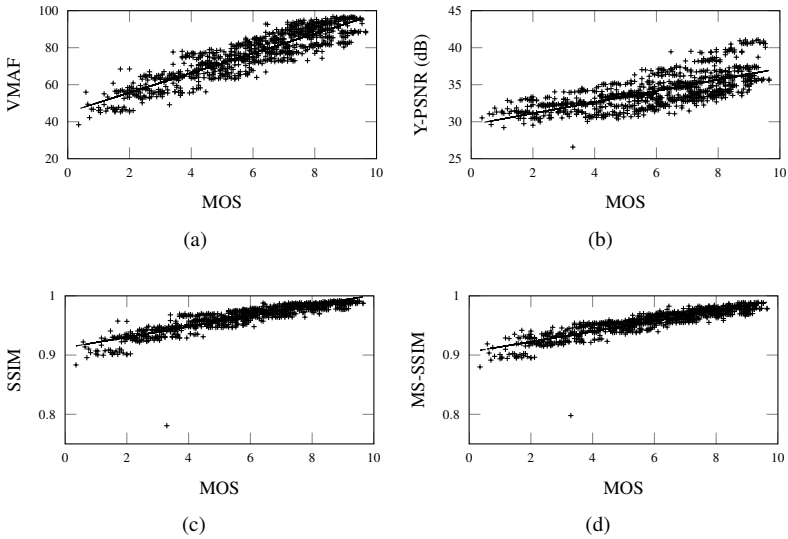

Figure 12.1: Acquired quality scores for the MPEG-JVET2018 dataset in relation to MOS, using (a) VMAF, (b) Y-PSNR, (c) SSIM, and (d) MS-SSIM.

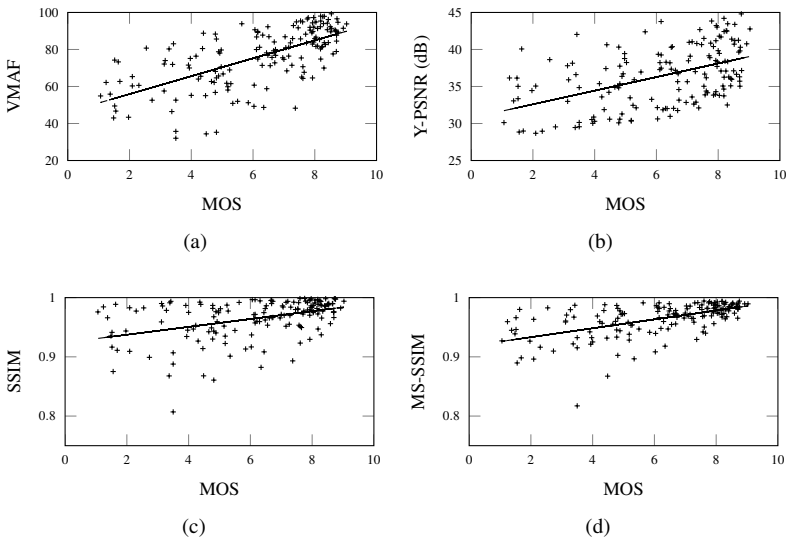

Figure 12.2: Acquired quality scores for the Tan et al. [38] dataset in relation to MOS, using (a) VMAF, (b) Y-PSNR, (c) SSIM, and (d) MS-SSIM.

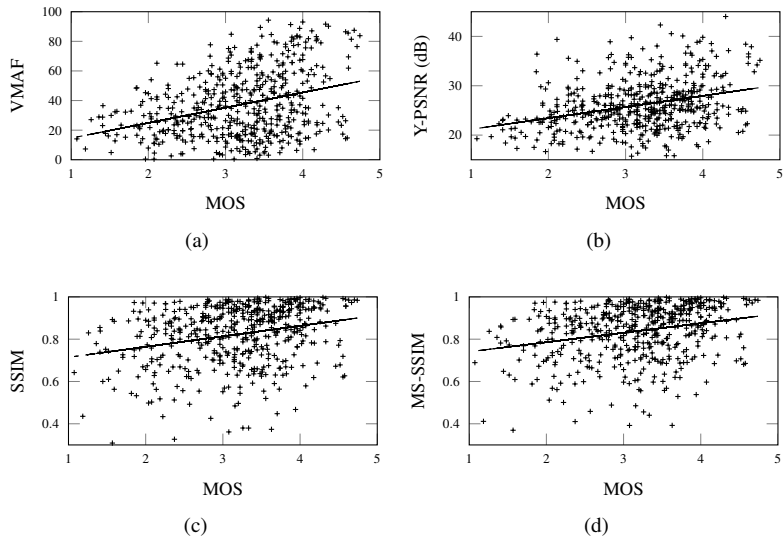

Figure 12.3: Acquired quality scores for the ITS4S dataset in relation to MOS, using (a) VMAF, (b) Y-PSNR, (c) SSIM, and (d) MS-SSIM.

SSIM, with MS-SSIM achieving the second best results. The very strong correlation of 0.90 achieved for the MPEG-JVET dataset shows that VMAF performs very well when the examined video sequences follow the default VMAF model's guidelines in terms of resolution, and when the impairments are related to coding impairments that VMAF was originally designed for. Similarly good results were achieved for the 1080p sequences of the Tan et al. dataset, which included video-coding-related impairments only, considering that the nominal correlation values for all metrics were lower than the ones achieved for the MPEG-JVET. While correlations to MOS ratings were significantly low for the ITS4S dataset, VMAF still achieved the highest correlation to viewer perceived quality. However, it is evident that the design constraints of VMAF, as well as of the other metrics, lead to significant differences between viewer-perceived quality scores and the computed scores. Our experimental evaluation showed that machine learning approaches for mapping objective video quality metrics to human-perceived video quality scores have great potential for providing perceptually accurate objective video quality metrics. However, more work is needed in order to provide metrics that would be suitable for a wide range of impairment types and would not be limited to specific usage scenarios.

References

[1] Barnett T, Jain S, Andra U, et al.. Cisco Visual Networking Index (VNI) Complete Forecast Update, 20172022; 2018. APJC Cisco Knowledge Network (CKN) Presentation.

[2] RFC 8216. HTTP Live Streaming; 2017. Apple Inc.
[3] ISO/IEC 23009-1. Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats; 2017.
[4] Aabed MA, AlRegib G. PeQASO: Perceptual Quality Assessment of Streamed Videos Using Optical Flow Features. IEEE Transactions on Broadcasting. 2019 Sep;65(3):534–545.
[5] Li Z, Aaron A, Katsavounidis I, et al.. Toward A Practical Perceptual Video Quality Metric; 2016. Accessed: 2019-11-08. https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652.
[6] Sheikh HR, Bovik AC. Image information and visual quality. IEEE Transactions on Image Processing. 2006 Feb;15(2):430–444.
[7] Zhou Wang, Bovik AC, Sheikh HR, et al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing. 2004 April;13(4):600–612.
[8] Chandler DM, Hemami SS. VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images. IEEE Transactions on Image Processing. 2007 Sep;16(9):2284–2298.
[9] Hemami SS, Reibman AR. No-reference image and video quality estimation: Applications and human-motivated design. Signal Processing: Image Communication. 2010;25(7):469 – 481. Special Issue on Image and Video Quality Assessment.
[10] Chikkerur S, Sundaram V, Reisslein M, et al. Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison. IEEE Transactions on Broadcasting. 2011 June;57(2):165–182.
[11] Murroni M, Rassool R, Song L, et al. Guest Editorial Special Issue on Quality of Experience for Advanced Broadcast Services. IEEE Transactions on Broadcasting. 2018 June;64(2):335–340.
[12] Winkler S, Mohandas P. The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics. IEEE Transactions on Broadcasting. 2008 Sep;54(3):660–668.
[13] Aabed MA, Kwon G, AlRegib G. Power of tempospatially unified spectral density for perceptual video quality assessment. In: 2017 IEEE International Conference on Multimedia and Expo (ICME); 2017. p. 1476–1481.
[14] Manasa K, Channappayya SS. An Optical Flow-Based Full Reference Video Quality Assessment Algorithm. IEEE Transactions on Image Processing. 2016 June;25(6):2480–2492.
[15] Seshadrinathan K, Bovik AC. Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos. IEEE Transactions on Image Processing. 2010 Feb;19(2):335–350.
[16] Ou Y, Ma Z, Liu T, et al. Perceptual Quality Assessment of Video Considering Both Frame Rate and Quantization Artifacts. IEEE Transactions on Circuits and Systems for Video Technology. 2011 March;21(3):286–298.
[17] Mok RKP, Chan EWW, Chang RKC. Measuring the quality of experience of HTTP video streaming. In: 12th IFIP/IEEE International Symposium on

Integrated Network Management (IM 2011) and Workshops; 2011. p. 485–492.

[18] Nauge M, Larabi M, Fernandez C. A reduced-reference metric based on the interest points in color images. In: 28th Picture Coding Symposium; 2010. p. 610–613.

[19] Tao S, Apostolopoulos J, Guerin R. Real-Time Monitoring of Video Quality in IP Networks. IEEE/ACM Transactions on Networking. 2008 Oct;16(5):1052–1065.

[20] Piamrat K, Viho C, Bonnin J, et al. Quality of Experience Measurements for Video Streaming over Wireless Networks. In: 2009 Sixth International Conference on Information Technology: New Generations; 2009. p. 1184–1189.

[21] Soundararajan R, Bovik AC. Video Quality Assessment by Reduced Reference Spatio-Temporal Entropic Differencing. IEEE Transactions on Circuits and Systems for Video Technology. 2013 April;23(4):684–694.

[22] Baik E, Pande A, Stover C, et al. Video acuity assessment in mobile devices. In: 2015 IEEE Conference on Computer Communications (INFOCOM); 2015. p. 1–9.

[23] Shahid M, Rossholm A, Lövström B, et al. No-reference image and video quality assessment: a classification and review of recent approaches. EURASIP Journal on Image and Video Processing. 2014 Aug;2014(1):40.

[24] Naccari M, Tagliasacchi M, Tubaro S. No-Reference Video Quality Monitoring for H.264/AVC Coded Video. IEEE Transactions on Multimedia. 2009 Aug;11(5):932–946.

[25] Wu Q, Li H, Meng F, et al. Toward a Blind Quality Metric for Temporally Distorted Streaming Video. IEEE Transactions on Broadcasting. 2018 June;64(2):367–378.

[26] Saad MA, Bovik AC, Charrier C. Blind Prediction of Natural Video Quality. IEEE Transactions on Image Processing. 2014 March;23(3):1352–1365.

[27] Mittal A, Saad MA, Bovik AC. A Completely Blind Video Integrity Oracle. IEEE Transactions on Image Processing. 2016 Jan;25(1):289–300.

[28] Li Y, Po L, Cheung C, et al. No-Reference Video Quality Assessment With 3D Shearlet Transform and Convolutional Neural Networks. IEEE Transactions on Circuits and Systems for Video Technology. 2016 June;26(6):1044–1057.

[29] Liu TJ, Lin YC, Lin W, et al. Visual quality assessment: recent developments, coding applications and future trends. APSIPA Transactions on Signal and Information Processing. 2013;2:e4.

[30] Lin JY, Liu T, Wu EC, et al. A fusion-based video quality assessment (fvqa) index. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific; 2014. p. 1–5.

[31] Li Z, Bampis C, Novak J, et al.. VMAF: The Journey Continues; 2018. Accessed: 2019-11-08. https://medium.com/netflix-techblog/vmaf-the-journey-continues-44b51ee9ed12.

[32] Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995 Sep;20(3):273–297.

[33] Li S, Zhang F, Ma L, et al. Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments. IEEE Transactions on Multimedia. 2011 Oct;13(5):935–949.

[34] Joint Video Exploration Team. Joint Exploration Model (JEM); 2019. Available from: https://jvet.hhi.fraunhofer.de/svn/svn_HMJEMSoftware/.

[35] Joint Video Exploration Team (JVET) of ITU-T VCEG (Q6/16) and ISO/IEC MPEG (JTC 1/SC 29/WG 11). Algorithm Description of Joint Exploration Test Model 7 (JEM7); 2017. Doc. JVET-G1001. Available from: http://phenix.it-sudparis.eu/jvet/.

[36] Methodology for the subjective assessment of the quality of television pictures. ITU-R; 2012. Recommendation ITU-R BT.500-13. Available from: https://www.itu.int/rec/R-REC-BT.500.

[37] Subjective video quality assessment methods for multimedia applications. ITU-T; 2008. ITU-T Rec. P.910.

[38] Tan TK, Weerakkody R, Mrak M, et al. Video Quality Evaluation Methodology and Verification Testing of HEVC Compression Performance. IEEE Transactions on Circuits and Systems for Video Technology. 2016 Jan;26(1):76–90.

[39] H.265: High efficiency video coding. ITU-T; 2016. ITU-T Rec. H.265.

[40] Parameter values for the HDTV standards for production and international programme exchange. ITU-R; 2002. ITU-R Rec. BT.709.

[41] JVT of ITU-T SG16/Q6 and ISO/IEC JTC1/SC29/WG11. AVC JM Reference Software Codebase, Version 18.5; 2013. Available from: http://iphome.hhi.de/suehring/tml/.

[42] JCT-VC of ITU-T SG16/Q6 and ISO/IEC JTC1/SC29/WG11. HEVC HM Reference Software Codebase, Version 12.1; 2013. Available from: http://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/.

[43] Pinson MH. ITS4S: A Video Quality Dataset with Four-Second Unrepeated Scenes; 2018. NTIA Technical Memo TM-18-532. Available from: https://www.its.bldrdoc.gov/publications/details.aspx?pub=3194.

[44] Pinson MH, Janowski L. AGH/NTIA: A Video Quality Subjective Test with Repeated Sequences; 2014. NTIA Technical Memo TM-14-505. Available from: https://www.its.bldrdoc.gov/publications/details.aspx?pub=2758.