On the accuracy of video quality measurement techniques

Deepthi Nandakumar
Amazon Video, Bangalore, India
nandakd@amazon.com

Yongjun Wu
Amazon Video, Seattle, USA
yongjuw@amazon.com

Hai Wei
Amazon Video, Seattle, USA
haiwei@amazon.com

Avisar Ten-Ami
Amazon Video, Seattle, USA,
avisarta@amazon.com

***Abstract*—With the massive growth of Internet video streaming, it is critical to accurately measure video quality subjectively and objectively, especially HD and UHD video which is bandwidth intensive. We summarize the creation of a database of 200 clips, with 20 unique sources tested across a variety of devices. By classifying the test videos into 2 distinct quality regions SD and HD, we show that the high correlation claimed by objective video quality metrics is led mostly by videos in the SD quality region. We perform detailed correlation analysis and statistical hypothesis testing of the HD subjective quality scores, and establish that the commonly used ACR methodology of subjective testing is unable to capture significant quality differences, leading to poor measurement accuracy for both subjective and objective metrics even on large-screen display devices.**

Keywords— *video quality, subjective testing methodology, objective video quality metrics, video streaming*

I. Introduction

By 2021, video is projected to touch 82% of global Internet traffic [1]. This introduces very specific challenges to streaming service providers, as they move large data chunks across the fragmented and unreliable Internet. One particular focus of optimization, is the accurate measurement of the end-user viewing experience, often referred to as Quality-of-Experience (QoE). The delivered video quality is a fundamental building block of the end-user experience metric, in addition to reliability and consistency metrics.

Measuring video quality reliably is a challenging problem. The ground truth dataset for video quality consists of the set of quality scores assigned by a group of human subjects to a large group of videos, as guided by a chosen subjective testing methodology [2]. However, as an iterative, in-loop optimization process, subjective testing is cumbersome, expensive and does not scale. Many objective metrics have been proposed to predict video quality, such as [3], [4], [5], [6], [7], [8] based on statistical signal processing, human visual system modelling and learning algorithms. The efficacy and prediction accuracy of these video quality metrics is usually evaluated using correlation coefficients (Pearson and Spearman's Rank) against ground truth scores on a test dataset. For maximum discrimination, subjective testing sessions typically show randomized videos, spanning the full quality range to the human subject. As a result, the efficacy of objective video quality metrics is also usually measured across the full quality range. In this study, we divide the full quality range into logical sections, and investigate the prediction accuracy of objective video quality metrics for individual quality sections. A very large fraction of the burgeoning video traffic is also guaranteed to be High-Definition (HD) and Ultra-High-Definition (UHD) video [1], and therefore measuring and optimizing for video quality in these high quality ranges is an imperative for streaming service providers. In this study, we lay specific emphasis on the measurement accuracy of subjective and objective video quality scores in this high quality range.

Globally, a healthy mix of devices with different screen sizes and form factors is projected to contribute to IP traffic in 2021 [1], ranging from smartphones (44%), to tablets (6%), PCs (19%) and TVs (24%). It is therefore necessary for streaming service providers to quantify the viewing experience based on the device, and possibly optimize the encoding and delivery process accordingly. As display technologies improve with OLED, plasma, HDR/WCG capable devices, it is important to ensure that the viewing quality experience is superlative and matched to customer expectations, while simultaneously optimizing for encoding and delivery efficiencies. In this study, we analyze the correlation of measured quality between different devices, again with special emphasis on the high quality range.

II. RELATED WORK

For improving video streaming quality, researchers in related fields have focused on developing reliable objective quality metrics, as well as studying the different influence factors and their impact on end-user Quality of Experience.

PSNR (Peak Signal to Noise Ratio) has been used as a default objective measure of video quality in the past. However, it has been shown that PSNR's correlation with the human visual experience is quite low [3]. Other full-reference metrics, such as Structural Similarity Index (SSIM) [4] and Multi-Scale Structural Similarity (MS-SSIM) [5], which extract and quantify the structural information present in the video have been proposed. VIF [8] measures the difference in the fidelity of the information conveyed between the reference and distorted video. In the context of Internet video streaming using adaptive HTTP streaming techniques, such as DASH [18] and HLS [19], a metric widely used in recent times is VMAF (Video Multi-method Assessment Fusion) [6] which focuses on measuring compression and scaling artifacts. A number of elementary features and corresponding ground truths are used to train a Support Vector Regressor (SVR) model. To measure the quality of a given video stream, the trained SVR predicts quality score for each video frame from elementary metrics and an aggregate score is computed.

For optimizing video streaming services, several studies on Quality of Experience (QoE) focus on qualifying and quantifying the subjective quality impact caused by playback device properties and viewing conditions. Redl e*t al.* conducted perceptual quality experiments comparing the effects of different HDTV devices including PC monitor, LCD TV and HD projector [13]. Catellier e*t al.* conducted exploratory experiments using five mobile devices in two testing environments [14]. Their study also performed a statistical analysis to investigate the influence of video resolution, viewing device and audio quality on perceived audiovisual quality. Furthermore, Li e*t al.* performed subjective experiments using a new Acceptance-Annoyance test methodology, aiming to quantify the perceptual difference due to user's device [5]. Their study focused on the comparison between Full HD TV and HD Tablet devices and also incorporated the service cost as an additional factor. All of these studies use videos that span the full quality range ranging from very poor to very high quality, and do not investigate the problem of accuracy over individual regions of the quality spectrum.

III. SUBJECTIVE EXPERIMENT CONFIGURATION

We conducted comprehensive subjective quality tests using a careful selection of videos that are representative of the videos present in the catalog of a top streaming service provider. Using the ITU-T P.913 Recommendation for assessment of Internet video [11], we conducted an Absolute Category Rating (ACR) test, on an 11-point scale. This experiment was representative of the streaming customer's viewing experience, with no double stimulus or side-by-side comparisons.

A. *Test Sequences*

Twenty full HD SRC test sequences were chosen from the catalog of a top streaming service provider, based on popularity and diversity of content. The diversity of chosen content can be represented by measurements of spatial and temporal information [23]. If the luma pixel values of the nth frame is represented as F_n, the Spatial Information (SI) and Temporal Information (TI) are defined as

$$SI = \max_n \{std_{i,j}[Sobel(F_n)]\}$$

$$TI = \max_n \{std_{i,j}[diff(F_n - F_{n-1})]\}$$

where Sobel (F_n) is the Sobel-filtered frame and diff ($F_n - F_{n-1}$) represents the pixel-wise difference of the frame from the previous frame. As shown in Fig 1, the source sequences span a wide range of the complexity space.

These were then encoded into 10 quality levels, with progressively increasing average bitrates and resolutions, encoded using the x265 v2.7 HEVC encoder [21]. The specific quality parameter used for encoding was CRF (Constant Rate Factor) that aims to achieve "uniform quality". CRF coupled with parameters defining Video Buffering Verifier (or decoding buffer) constraints were used to encode these video sources. We note that this creates a spread of bitrates for each quality level, and the mean bitrates for each of these quality levels are as noted below in TABLE I.

These 10 quality levels were further classified into 2 quality regions. The first quality region ("SD") consisted of 6 quality levels, encoded at 4 different frame sizes, where the frame sizes were lower than 1280x720 (both width and height lesser than 1280 and 720 respectively).

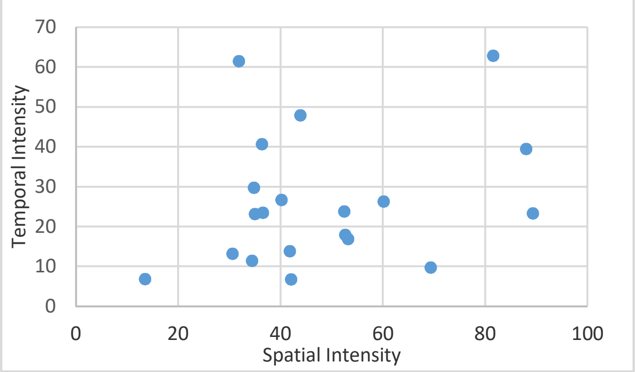

Fig 1: Spatial intensity plotted against temporal intensity for the 20 sources used in the subjective test.

The second quality region ("HD") consisted of videos encoded at 2 different frame sizes, with width or height greater than or equal to 1280 and 720 respectively, and less than or equal to 1920 and 1080 respectively. We avoid prescribing exact frame sizes for each quality level, since this is decided based on the frame size of the SRC video and its display aspect ratio. In all, there were 200 PVS and 20 SRC sequences.

B. *Viewing Devices and Test Environment*

The tests were conducted in a laboratory area with low ambient lighting (just enough to fill out scoring sheets), with non-reflecting black or grey surfaces. We used 3 viewing devices to conduct our subjective tests. The first device was a UHD-TV, a 65 inch OLED commercial TV set, with 3840x2160 resolution. 3 subjects were seated at a distance of 2H, with a viewing angle less than 45degree to the central screen axis. The second device was a PC monitor, 23 inch commercial monitor, with 1920x1080 resolution. 2 subjects were seated at a distance of 2H, with a viewing angle less than 30degree to the central screen axis.

TABLE I. DESCRIPTION OF PVS

Quality Level Index	Quality Region	Mean Bitrate (kbps)
1	SD	54 kbps
2	SD	75 kbps
3	SD	128 kbps
4	SD	252 kbps
5	SD	520 kbps
6	SD	700 kbps
7	HD	1120 kbps
8	HD	1961 kbps
9	HD	2415 kbps
10	HD	4950 kbps

TABLE II. CORRELATION ANALYSIS OF MOS (FULL DATASET) WITH OBJECTIVE METRICS

	Pearson's Correlation Coefficient					Spearman's Rank Correlation Coefficient				
	VMAF	PSNR	SSIM	MS-SSIM	VIF	VMAF	PSNR	SSIM	MS-SSIM	VIF
Tablet	0.9048	0.7773	0.8209	0.8127	0.8639	0.8563	0.6924	0.8118	0.7745	0.8358
Monitor	0.9145	0.7788	0.8346	0.8293	0.8815	0.8813	0.7005	0.8396	0.7991	0.8533
UHD-TV	0.9045	0.7652	0.8151	0.8099	0.8650	0.8875	0.6960	0.8326	0.7857	0.8599

The third device was a 9.7 inch tablet with an AMOLED screen, with a resolution of 2048x1536. Viewers, one per tablet, were seated in front of the device, which was mounted at an angle of 80degree to the desk plane.

C. *Subjects and Training*

For each viewing device, 10 sessions were conducted with 24 subjects. All the subjects were tested for visual acuity (normal or corrected to normal) and color perception. The subjects were asked to rate the quality of video, considering themselves paying subscribers of an Internet streaming service. The 11-point numerical scale was described to them as shown in TABLE III. The scores were explained clearly to people, and they were given time to practice and internalize the scoring procedure. Different subjects watched the videos in at least 6 different ordering sequences. Outlier analysis and screening of observers was performed in accordance with the rules specified in [10].

TABLE III. DESCRIPTION OF SCORING SCALE USED FOR SUBJECTIVE EXPERIMENTS

Score	Quality Description	
10	Perfect	
9	Showing slight artefacts	somewhere
8		everywhere
7	Showing artefacts	somewhere
6		everywhere
5	Clearly showing artefacts	somewhere
4		everywhere
3	Showing annoying artefacts	somewhere
2		everywhere
1	Severely affected by artefacts	somewhere
0		everywhere

IV. CORRELATION WITH OBJECTIVE METRICS

In this section, we discuss the correlation of MOS scores on each device with objective video quality metrics and with each other, using Pearson Linear Correlation (PLCC) and Spearman Rank Order Correlation (SROCC) metrics. The objective quality metrics chosen were VMAF, PSNR, SSIM, MS-SSIM and VIF. All the metrics were computed at the original source resolution, after bicubic upscaling of the encoded PVS representations back to source resolution. TABLE IV shows that the overall MOS scores across devices are well-correlated, with slightly higher agreement between the HD monitor and UHD TV, than with the tablet device. This is reasonable since the viewing experience of the tablet device is very different from that of the HD monitor and UHD TV.

TABLE IV: CORRELATION ANALYSIS OF MOS ACROSS DEVICES

	PLCC			SROCC		
	Tablet	Monitor	UHD-TV	Tablet	Monitor	UHD-TV
Tablet	1.000	0.945	0.952	1.000	0.898	0.912
Monitor	0.945	1.000	0.971	0.898	1.000	0.946
UHD-TV	0.952	0.971	1.000	0.912	0.946	1.000

Among objective metrics, VMAF has the highest correlation with MOS scores for all devices, followed by VIF, SSIM, MS-SSIM and PSNR, as shown in TABLE II. The 95% confidence interval for Pearson correlation between MOS and VMAF for the HD monitor, calculated using the Fisher's z-transformation as a function of the sample size (*n=200*) is +/- 0.0071. Also, there is a strong linear relationship between MOS and VMAF – the adjusted R-square values (or explained variance) for the Tablet, HD-Monitor and UHD-TV are 0.8365, 0.8362 and 0.8182 respectively. The resolving power of VMAF for these devices, as calculated at the 95% and 75% confidence level are 28.209 and 7.62 for the tablet device, 22.79 and 6.62 for the monitor and 22.468 and 6.38 for the UHD-TV respectively, indicating the required delta at which viewers are statistically likely to notice a quality difference. As with the correlation metrics, there is higher agreement between the monitor and UHD-TV, than with the tablet device. The scatter plot showing the relationship between MOS and VMAF for all devices is shown in **Error! Reference source not found.**2.

V. ANALYSIS BASED ON QUALITY REGIONS

The results shown in TABLE II above appear to be consistent with previous analysis performed on the correlation of objective metrics with Mean Opinion Scores [20]. Our objective was to understand and measure the effectiveness of objective metrics depending on the underlying "quality region" they fall under. This is indirectly inspired by insights such as the thresholding effect of consumer satisfaction, and other subjective testing methodologies, such as Acceptance-Annoyance (AccAnn) proposed in [16], where subjects characterized video quality into acceptable/annoying. The "quality regions" that we divided out videos into, were "SD" and "HD", based on frame size, as explained in Section 3. The reasoning behind this division was partly psychological, since viewers have a higher degree of satisfaction in video quality when the video is confirmed to be HD (such as a screen indicator).

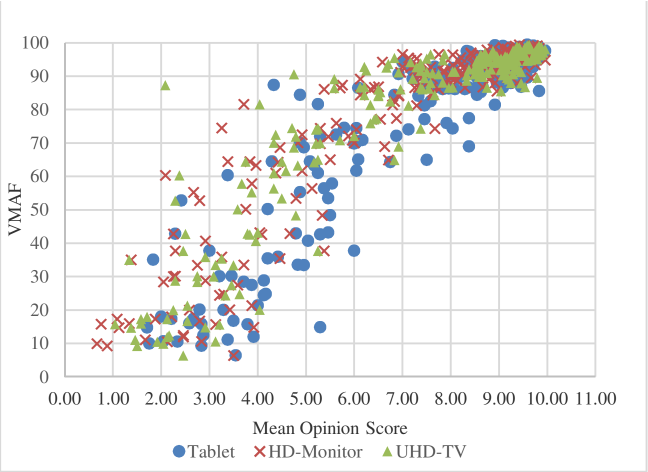

Fig 2: Scatter plot between MOS and VMAF (full quality region)

While our subjective viewing experiment did not include any explicit notifications of resolution, we confirm this hypotheses in TABLE V, which shows the results of a T-test between the Gaussian distributions centered at the MOS values with a known standard deviation. A value of '0' indicates that the highest SD quality level was statistically inferior (worse video quality) than the lowest HD quality level, a value of '1' indicates that the highest SD quality level was statistically superior to the lowest HD quality level, and a value of '_' indicates the 2 quality levels were statistically equivalent. Each sub-entry corresponds to the 20 reference videos used in the study.

TABLE V shows that the increase in MOS scores between the highest quality SD video and the lowest quality HD video are consistently and statistically significant for the UHD-TV and HD monitor at the 95% confidence level, and less so for the tablet device. We also note the appreciable and consistent increase in bitrate between these two quality levels as described in TABLE I. For simplicity of analysis, we use the same definition of SD and HD quality regions for all devices.

A. *Correlation with objective metrics in SD quality region*

In this section, we segregate the MOS scores that fall in the SD region (6 points per SRC clip, total 120 test videos). The correlation values for all PVS that fall in SD region are shown in TABLE VI, and we note that the SD correlation scores are slightly lower than, but still very close to the PLCC and SROCC values for the full set of test videos in TABLE II. The 95% confidence interval for Pearson correlation between MOS and VMAF for the HD monitor on SD videos, calculated using the Fisher's z-transformation as a function of the sample size *(n=120)* is +/- 0.018. The strength of the linear relationship between MOS and VMAF, explained by adjusted R-square values (or explained variance) for Tablet, HD-Monitor and UHD-TV are 0.781, 0.7976 and 0.7867 respectively. Again, this is slightly lower than, but still very close to the R-square values for the overall set of MOS scores.

B. *Correlation with objective metrics in HD quality region*

In this section, we analyze the correlation of MOS scores that fall in the HD region (4 points per clip, total 80 test videos). The correlation scores between MOS and all objective metrics for all PVS that fall in the HD region are shown in TABLE VII, and are significantly lower than the overall MOS scores in TABLE II as well as the MOS scores for SD test videos in TABLE VI.

TABLE V. RESULTS OF T-TEST ON SUBJECTIVE SCORES BETWEEN THE HIGHEST SD QUALITY LEVEL AND THE LOWEST HD QUALITY LEVEL. EACH SUB-ENTRY CORRESPONDS TO 20 SRC VIDEOS.

	Lowest HD Quality Level		
	Tablet	HD-Monitor	UHD-TV
Highest SD Quality Level	_00000_ _01 10_001_000	0000_00000 _00000000_	0000000000 0000_00000

The 95% confidence interval for Pearson correlation between MOS and VMAF for the HD monitor on HD videos, based on sample size *(n=80)* is +/- 0.093. The adjusted R-square values are also very low, at 0.067, 0.1792 and 0.17 respectively for the tablet, monitor and UHD-TV. Fig 3 shows the scatter plot for HD videos between VMAF and MOS, and we can see that the linear relationship is a lot weaker than for the overall set of scores. The plot also suggests a heteroscedastic relation between VMAF and MOS. Box-Cox transformation of the VMAF values did not yield any meaningful increases in correlation metrics.

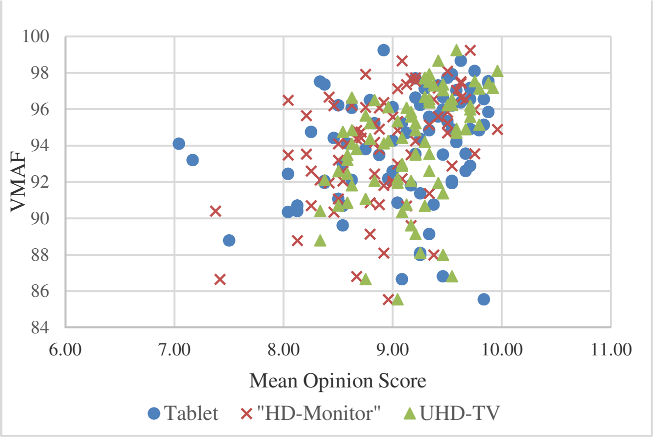

Fig 3: Scatter plot between MOS and VMAF (HD quality region)

To better understand the phenomenon of poor correlation exhibited by MOS and all objective quality metrics, we further examine the statistical distribution of the MOS scores in the HD region itself. Our objective is to understand if the poor correlation is a result of inaccurate modeling of the objective metric, or whether this is caused due to insufficiencies in the ground truths for HD subjective quality scores. The 4 HD quality levels include 2 levels of 720p resolution encoded using variable bitrate encoding (with average bitrates over 20 clips of 1.1 Mbps and 1.9 Mbps), and 2 levels of 1080p resolution (with average bitrates over 20 clips of 2.4 Mbps and 4.8 Mbps). Our objective was to evaluate whether subjects are able to adequately distinguish between all 4 HD quality levels with a significant bitrate span between 1.1 Mbps and 4.8 Mbps with any degree of predictability for 20 different sources. To this end, TABLE IX shows the results of the t-test for significance, comparing the statistical significance between the two 720p quality levels and the two 1080p quality levels.

TABLE VI. CORRELATION ANALYSIS OF MOS (SD REGION) WITH OBJECTIVE METRICS

	Pearson's Correlation Coefficient					Spearman's Rank Correlation Coefficient				
	VMAF	PSNR	SSIM	MS-SSIM	VIF	VMAF	PSNR	SSIM	MS-SSIM	VIF
Tablet	0.8688	0.7637	0.7844	0.7712	0.8242	0.8193	0.7001	0.7468	0.7410	0.7555
Monitor	0.8931	0.7765	0.8244	0.8142	0.8679	0.8244	0.7051	0.7700	0.7542	0.7752
UHD-TV	0.8870	0.7629	0.8071	0.7972	0.8533	0.8260	0.6983	0.7578	0.7423	0.7658

TABLE VII. CORRELATION ANALYSIS OF MOS (HD REGION) WITH OBJECTIVE METRICS

	Pearson's Correlation Coefficient					Spearman's Rank Correlation Coefficient				
	VMAF	PSNR	SSIM	MS-SSIM	VIF	VMAF	PSNR	SSIM	MS-SSIM	VIF
Tablet	0.2599	0.1088	0.1668	0.0588	0.1731	0.3243	0.0723	0.2095	0.1118	0.2294
Monitor	0.4233	0.0933	0.2719	0.1628	0.3304	0.4417	0.1138	0.2708	0.1797	0.1698
UHD-TV	0.4124	0.0876	0.1876	0.0578	0.1807	0.4794	0.0936	0.1967	0.0930	0.2173

TABLE VIII. RESULTS OF T-TEST BETWEEN 4 HD QUALITY LEVELS ON THE UHD-TV. ENTRIES ABOVE THE MAIN DIAGONAL ARE EXACT INVERSES OF THOSE BELOW THE MAIN DIAGONAL. EACH SUB-ENTRY CORRESPONDS TO 20 SRC VIDEOS.

	Quality Level 1	Quality Level 2	Quality Level 3	Quality Level 4
Quality Level 1 (720p) 1.1 Mbps	_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _	0_000_ _000 00_0_00000	00_0_0000_ _0_ _ _0_ _0_	0_0_ _ 00000 _000000_0_
Quality Level 2 (720p) 1.9 Mbps	1_111_ _111 11_1_11111	_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _	0_0010_ _0_ _01_ _0_00_	0_0_ _ 0_ _00 _0_0000_0_
Quality Level 3 (1080p) 2.4 Mbps	11_1_1111_ _1_ _ _1_ _1_	1_1101_ _1_ _10_ _1_11_	_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _	_ _ _1_ _ _ _ _0 _ _ 0_0_0_ _ _
Quality Level 4 (1080p) 4.8 Mbps	1_1_ _11111 _111111_1_	1_1_ _1_ _11 _1_1111_1_	_ _ _0_ _ _ _ _1 _ _1_1_1_ _ _	_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

A value of '0' indicates that the row (720p quality levels) is statistically inferior (worse video quality) to the column (1080p quality levels), a value of '1' indicates that the row (720p quality levels) is statistically superior to the column (1080p quality levels), and a value of '_' indicates that the 1080p and 720p quality levels are statistically equivalent. Each sub-entry corresponds to the 20 reference videos used in the study.

The results in TABLE IX show that for a significant proportion of the videos, viewers are unable to statistically distinguish between the significant changes that should impact perceived quality – such as a change in resolution from 720p to 1080p, and a change in bitrate from 1.1 to 4.8 Mbps. Out of 20 videos, we see that the subjective scores of 8, 9 and 7 videos (corresponding to the tablet, monitor and UHD-TV respectively) do not statistically distinguish between the 720p quality levels and the 1080p quality levels with 95% confidence (indeed, in some cases the 720p quality levels are statistically superior than the 1080p levels). Using TABLE V as contrast, we can conclude that human subjects can reliably appreciate a change in resolution from "SD" to "HD", particularly on an HD monitor and UHD-TV for most content types, however, they can less reliably differentiate between 720p and 1080p.

TABLE IX. RESULTS OF T-TEST ON SUBJECTIVE SCORES BETWEEN 720P AND 1080P QUALITY LEVELS.

	1080p Quality Levels		
	Tablet	HD-Monitor	UHD-TV
720p Quality Levels	_10_1_0000 0000_00101	0_0_010_0_ 000_ _0_ _00	0_0_100_00 _0_000000_

We now use the subjective quality scores on the UHD-TV to dive deeper into statistical differentiation within the individual HD quality levels, based on the straightforward assumption that the UHD-TV is better positioned to display differences in HD quality than other devices. TABLE VIII shows that as the quality levels increase, the number of titles in which the subjective scores are statistically similar to the previous quality level with a confidence level of 95%, increases. For instance, the subjective scores for Quality Level 4 (1080p 4.8 Mbps average bitrate) are higher than Quality Level 3 (1080p 2.4 Mbps average bitrate) only for 4 out of the 20 videos, with a confidence of 95%. The difference between Quality Level 1 (720p 1.1 Mbps) and Quality Level 4 (1080p 4.8 Mbps) appear meaningful (14 out of 20 videos show statistically higher scores), however even this is lower than anticipated, since the display device is a UHD-TV, we expected that quality differences between these levels would be consistent and pronounced. TABLE V, TABLE VIII and TABLE IX together demonstrate the underlying problem of poor correlation of MOS scores with objective quality metrics. The ACR method of subjective testing does not enable subjects to accurately distinguish between significant quality changes (resolution changes from 720p to 1080p, or bitrate changes ranging from 1.1 Mbps to 4.8 Mbps), even on a large screen device like the 65 inch UHD-TV. Videos belonging to this quality range do not possess obvious artifacts, and hence the power of ACR testing to distinguish between significantly different quality changes is poor. Note that during the ACR test sessions, videos over the full quality range (SD and HD) were shown, and there was no subject fatigue due to the similarity of video quality.

This also explains the poor correlation exhibited by MOS and objective quality metrics such as VMAF, which are learned algorithms, generated from ground truth based on ACR-HR (Absolute Category Rating with Hidden Reference), which is statistically very similar to ACR. The bias manifested in the ground truth data for this application, is similar to the imbalance manifested in machine learning applications due to overlapping of scores, which could be fixed using a combination of over-sampling and data-cleaning methods [22]. In addition, we also plan to investigate the use of subjective testing methodologies with more discriminative power such as viewing tests with expert subjects and/or paired comparison tests to explore whether this leads to better ground truth data distributions in the high quality range.

VI. CONCLUSIONS AND FUTURE WORK

We summarized the creation of a dataset that tested clips encoded over a large quality range, on a variety of devices. Correlation analysis showed that while subjective and objective quality metrics are correlated very well in the lower SD quality range, correlation in the HD quality range, in the absence of obvious artifacts, is significantly lower. Detailed significance testing shows that the subjective scores generated using ACR testing do not consistently distinguish between quality levels in the HD region, even when significant quality differences are expected. This also explains the poor correlation with top-performing objective metrics like VMAF, since these learning metrics have been trained on subjective datasets with a similar bias. Thus, we conclude that existing methodologies for both subjective and objective measurement of video quality in the HD region are insufficient and need significant improvement, in keeping with the leaps in display technologies and device resolutions. In future works, we plan to extend our analyses to different subjective testing methodologies with higher discriminative power in the high quality range, and train objective metrics using more discriminative subjective datasets. We also plan to extend our tests to more display devices and UHD encoded videos as well.

REFERENCES

[1] Cisco Visual Networking Index: Forecast and Trends, 2017-2022, https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

[2] M.H. Pinson,S. Wolf, “Comparing subjective testing methodologies”, Proceedings of the SPIE, vol. 5150, pp. 573-582, 2003.

[3] Z. Wang, A.C. Bovik, “Mean Squared Error: Love it or leave it?”, IEEE Signal Processing Magazine, vol. 26, no. 1, pp. 98-117, 2009.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", IEEE Transactions on Image Processing, vol. 13, no.4, pp. 600-612, 2004.

[5] Z. Wang, E.P. Simoncelli, A.C. Bovik, “Multi-scale structural similarity for image quality assessment”, 37th IEEE Asilomar Conference on Signals, Systems and Computers, 2003

[6] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, “Toward a practical perceptual video quality metric.” http: //techblog.netflix.com/2016/06/toward-practical-perceptual-video.html.

[7] S. Li, F. Zhang, L. Ma, and K. N. Ngan, “Image quality assessment by separately evaluating detail losses and additive impairments,” IEEE Trans. Multimedia, vol. 13, no. 5, pp. 935–949, 2011.

[8] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” IEEE Trans. Image Process., vol. 15, no. 2, pp. 430–444, 2006.

[9] C. Bampis, A.C. Bovik, Z. Li, “A Simple Prediction Fusion Improves Data-driven Full-Reference Video Quality Assessment Models”, Picture Coding Symposium, pp. 298-302, 2018.

[10] ITU-R BT.500-13, “Methodology for the subjective assessment of the quality of television pictures”, International Telecommunication Union, Geneva, Switzerland, 2012.

[11] ITU-T P.913, “Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment”, International Telecommunication Union, Geneva, Switzerland, 2016.

[12] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, “Aquality-of-experience index for streaming video,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 1, pp. 154–166, 2017.

[13] A. Redl, C. Keimel, and K. Diepold, “Influence of viewing device and soundtrack in hdtv on subjective video quality,” in Image Quality and System Performance IX, vol. 8293, 2012.

[14] A. Catellier, M. Pinson, W. Ingram, and A. Webster, “Impact of mobile devices and usage location on perceived multimedia quality,” in Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on. IEEE, 2012, pp. 39–44.

[15] N. O. Sidaty, M.-C. Larabi, and A. Saadane, “Influence of video resolution, viewing device and audio quality on perceived multimedia quality for steaming applications,” in Visual Information Processing (EUVIP), 2014 5th European Workshop on. IEEE, 2014, pp. 1–6.

[16] J. Li, L. Krasula, Z. Li, Y. Baveye, P. L. Callet, “Quantifying the influence of devices on Quality of Experience for Video streaming”, PCS 2018.

[17] M.H. Pinson, S.Wolf, “Techniques for evaluating objective video quality models using overlapping subjective data sets”, NTIA Technical Report, TR-09-457, 2008.

[18] ISO/IEC 23009-1:2014, Dynamic Adaptive Streaming over HTTP (DASH), https://www.iso.org/standard/65274.html

[19] HTTP Live Streaming 2nd Edition, https://tools.ietf.org/html/draft-pantos-hls-rfc8216bis-03

[20] Reza Rassool, “VMAF Reproducibility: Validating a perceptual practical video quality metric”, IEEE International Symposium on Broadband Multimedia Systems and Braodcasting, July 2017

[21] http://x265.org, H.265 Video Codec

[22] G. E. A. P. A. Batista, R. C. Prati, M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data”, ACM SIGKDD Explor. Newsl. 6, 1 (June 2004)

[23] ITU-T P.910, “Subjective Video Quality Assessment Methods for Multimedia Applications”, 2008