Characterization of the HEVC Coding Efficiency Advance Using 20 Scenes, ITU-T Rec. P.913 Compliant Subjective Methods, VQM, and PSNR

Andrew Catellier, Margaret Pinson
Institute for Telecommunication Sciences
NTIA, U.S. Department of Commerce
Boulder, Colorado, USA
{*acatellier, mpinson*}*@its.bldrdoc.gov*

***Abstract*—The new video coding standard, MPEG-H Part 2 High Efficiency Video Coding (HEVC) or H.265, was developed to be roughly twice as efficient as H.264/AVC—meaning H.265/HEVC could deliver the same quality as H.264/AVC using roughly half the bitrate. In this paper we describe a subjective experiment designed to test this claim. We present an experiment using 20 different 1080p 29.97 fps scenes and 12 impairment levels spanning MPEG-2, H.264/AVC and H.265/HEVC. Additionally we compare the results obtained from the subjective assessment to quality estimates from two objective metrics: VQM and PSNR. Our subjective results show that H.265/HEVC can deliver the same quality at half the bitrate compared to H.264/AVC and can perform better at one quarter the bitrate compared to MPEG-2 in many, but not all, situations. For all 20 scenes coded with H.265/HEVC at 4 Mbps mean opinion scores span 38% of the subjective scale, which indicates the importance of scene selection. Objective quality estimations of HEVC have a low correlation with subjective results (0.60 for VQM, 0.64 for PSNR).**

***Keywords*-video coding; compression; subjective testing; AVC; HEVC; MPEG2; h.264; h.265; VQM; PSNR**

I. Introduction

The video coding standard H.264/AVC (also known as MPEG-4 Part 10 Advanced Video Coding) [1] is very widely deployed and has been used to compress countless hours of video. It was designed to replace the previous MPEG/ISO joint standard MPEG-2 [2], first released in 1995. Though annexes for H.264/AVC standard were developed until 2009, the first version of the standard was completed in 2003. H.264/AVC was reported to be twice as efficient as MPEG-2, where efficiency is defined as the bitrate required to deliver a fixed visual quality level. Work began in 2004 to once again double video coding efficiency in what would eventually become the successor to H.264/AVC. In January of 2013 H.265/HEVC (also known as MPEG-H Part 2 High Efficiency Video Coding) [3] received final draft status and the race to build commercial implementations intensified. Since then some level of HEVC support has been added to three major computing platforms (iOS, Android, and Windows 10), the Blu-Ray Disc Association has announced that HEVC will be used for 4K Blu-Ray discs, and hardware manufacturers have released hardware decoders.

The benefits of using a video coding standard twice as efficient as the stalwart H.264/AVC (and four times as efficient as MPEG-2) in a spectrum-hungry society are obvious. Some legacy cable systems still use MPEG-2 for video distribution. Updating to a more modern codec could increase the number of channels it's possible to distribute or increase delivered video quality. As more and more video entertainment is delivered using IP networks, the importance of high efficiency video coding increases. However, it is also important to validate codec efficiency claims in order to prevent unintended consequences like a reduction in delivered video quality for certain content types.

In this work we build on the research presented in [4] to design, implement, and conduct a subjective video quality test to measure the efficiency of H.265/HEVC compared to its predecessors. We introduced coding distortions by artificially controlling coding rate in order to compare the efficiency of MPEG-2, H.264/AVC and H.265/HEVC. Additionally we provide some objective video quality estimates using PSNR and VQM [5] to allow for comparison of subjective and objective methods.

First, in Section II, we survey the research conducted concerning HEVC efficiency. In Section III we discuss the design of the experiment we conducted, our source video selection, stimulus preparation, exactly how the subjective assessment was conducted, and the implementation of an objective assessment. Our results are shown in Section IV and we conclude our paper in Section V.

II. Prior Work

A variety of papers have compared H.265 and H.264 using peak signal to noise ratio (PSNR). These papers are not cited because PSNR is much less accurate than subjective testing; the accuracy of such analyses relies upon the unproven reliability of PSNR for accurate quality estimates when comparing these two codecs. We instead focus on prior subjective tests.

This paper was presented at and will appear in the Proceedings of the IEEE International Symposium on Multimedia, Miami, FL, December 14-16, 2015.

Pinson et al. [4] compared the coding efficiency of MPEG-2 and H.264 using commercial grade software encoders. This subjective test used ITU-T Rec. P.910, the absolute category rating (ACR) method, 12 video sequences, 1080i 30fps video, a $49''$ monitor, and 24 viewers. Ignoring packet loss impairments, MPEG-2 was impaired at four levels (6, 8.5, 12.5, and 18 megabits per second or Mbps) and H.264 was impaired at five levels (2, 3.5, 6, 10, and 17 Mbps). Pinson et al. [4] found an overall isoquality bit-rate reduction of $\approx$50%, with diminished advantages at high bit-rates ($\geq$17 Mbps) as measured quality for both coders converged.

Ohm et al. [6] presents a preliminary subjective test that compares the reference implementations of H.265 and H.264. This experiment used ITU-R Rec. BT.500, the double stimulus impairment scale (DSIS), nine video sequences, four impairment levels for each codec, a $50''$ monitor, and 24 viewers. The nine source videos were in 1080p format and span a variety of frame rates (24, 30, 50 and 60 fps). Ohm et al. [6] measures an overall isoquality bit-rate reduction of 49.3%. These results are supported by Weerakkody et al. [7], who perform a similar experiment on ultra high definition video (UHD) using a $56''$ monitor.

Garcia and Kalva [8] compare the reference implementations of the H.264 and H.265 in a low bandwidth mobile environment. This subjective test used ITU-R Rec. BT.500, DSIS, eight video sequences, two impairment levels, a $4.3''$ monitor, and 25 viewers. Each video was encoded with whichever quantization parameter (QP) value yielded a bit-rate closest to 400 Kbps and 200 Kbps. The videos were transmitted at 640×360 resolution yet displayed at 480×272 resolution. Garcia and Kalva concluded H.264 and H.265 yielded similar quality for these low bit-rates.

The authors of Ohm et al. [6] are active in MPEG and their subjective test is designed from a video coding algorithm development perspective. Each scene contains visual characteristics that exercise specific aspects of the codec (e.g., motion estimation). Consequently, most of the video sequences have no scene cuts. Second, impairment levels are specified by QP values, so each impairment level produces video clips with roughly similar visual quality but dissimilar bit-rates. The subjective data proved quality equivalence for pairs of encoded videos, but video bitrates were not explicitly chosen.

In this work we take an application-based approach by using scenes of many different content types with varying levels of complexity and varying numbers of scene cuts. We also specify coding rates as one may be required to do when transmitting a video sequence. Finally, we use the absolute category rating (ACR) method and the newly approved ITU-T Rec. P.913 subjective testing methodology.

III. Experiment Design

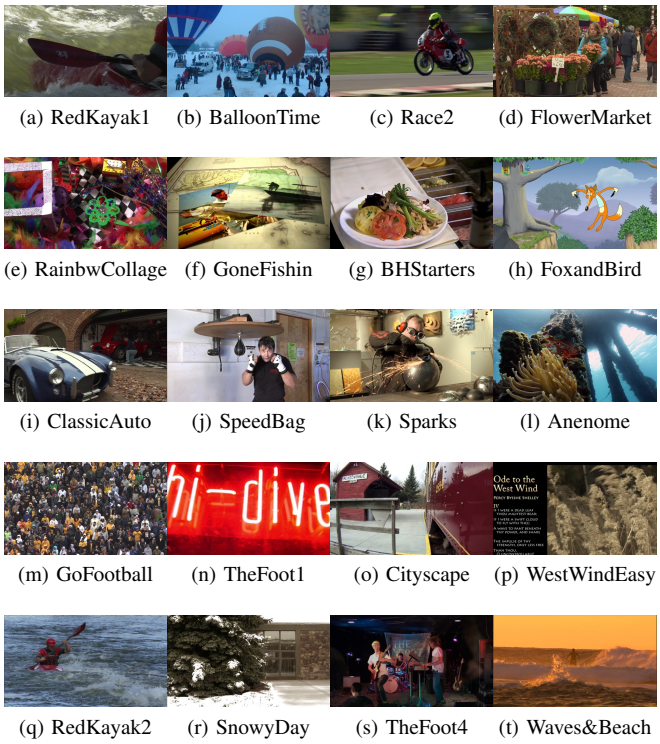
In order to vet efficiency claims we use anchor points that are well-understood in the video coding community. In [4] MPEG-2 was found to have good quality when processing 1080i 30fps video at 18 Mbps and was quickly approaching poor quality at 6 Mbps. H.264 was found to have good quality at 10 Mbps and poor quality at 2 Mbps. These anchor points are the basis of the test design shown in Table I. Processing video using MPEG-2 at 4, 8, and 16 Mbps should result in mean opinion scores ranging from 2 to 4 and allows for comparison of H.264 at half of each bitrate and H.265 at one quarter of each bitrate. Therefore the design calls for testing H.264 at 8, 4 and 2 Mbps and H.265 at 4, 2 and 1 Mbps. We include H.264 at 16 Mbps and H.265 at 16 and 8 Mbps in order to investigate where mean opinion score (MOS) saturates towards the top of the scale. This results in $3 + 4 + 5 = 12$ impairment levels.

A. Source Video

As explained in [9], selecting video clips for use in subjective video quality tests requires careful consideration. Content must not be offensive, controversial, polarizing, or distracting. Camerawork and editing should be high quality and the sequences used during a test should span a wide range of coding complexity. The number of clips to be used should be balanced compared with the number of impairments in an attempt to mitigate viewer fatigue. We therefore chose to use 20 source sequences. Figure 1 shows frames from the 20 video clips chosen for this test. These video clips are available for download at www.cdvl.org [10].

Additionally, [9] explains the importance of using a large number of source sequences to more fully exercise a system under test. Testing 20 source video sequences with 12 impairment levels plus the original would result in $20 \times 13 = 260$ processed video sequences (PVS). This would result in a very long test (around 2 hours and 20 minutes including four 15 minute breaks) so it was necessary to reduce the scope of testing. To do this, we grouped the 20 source sequences into four sets of five videos.

The sequences were visually examined and manually classified by coding complexity (high, medium, and low) by an expert viewer. The first set of five source sequences, or the *all impairments* set, was chosen such that the set contained sequences with high to medium coding complexities and would therefore be challenging to the codecs at all tested bitrates. The next set of five source sequences, the *low bandwidth* set, was chosen such that the set contained source sequences with complexity levels that would be challenging to the codecs at lower bitrates. These lower-complexity source sequences didn't seem to challenge the codecs at higher bitrates; preliminary viewing revealed that few coding artifacts were visible. Similarly, the *medium* and *high bandwidth* sets were chosen such that the sets contained source sequences with complexity levels that would be challenging to the codecs at medium to high bitrates. In this way we reduced the number of PVSs in our test while maximizing coverage of the problem space.



(a) RedKayak1 (b) BalloonTime (c) Race2 (d) FlowerMarket

(e) RainbwCollage (f) GoneFishin (g) BHStarters (h) FoxandBird

(i) ClassicAuto (j) SpeedBag (k) Sparks (l) Anenome

(m) GoFootball (n) TheFoot1 (o) Cityscape (p) WestWindEasy

(q) RedKayak2 (r) SnowyDay (s) TheFoot4 (t) Waves&Beach

Figure 1. Video clip screen captures. Subfigures 1a–1e are the all impairments set, 1f–1j are the low bandwidth set, 1k–1o are the medium bandwidth set, 1p–1t are the high bandwidth set.

Table I
HRC TABLE. BITRATES ARE IN MEGABITS PER SECOND, A BULLET INDICATES INCLUSION IN THE SUBJECTIVE TEST.

set	MPEG-2			H.264				H.265				
	4	8	16	2	4	8	16	1	2	4	8	16
all	•	•	•	•	•	•	•	•	•	•	•	•
low	•			•	•			•	•	•		
med		•			•	•			•	•	•	
high			•			•	•			•	•	•

B. Stimulus Preparation

From the set of all PVSs, the subset used for the subjective assessment is shown in Table I. For example, each PVS of the video clip "RedKayak1" was included in the subjective assessment. This means that $3 + 4 + 5 = 12$ impaired versions and the original video clip were tested, because this clip was in the *all impairments* set. Further, the subjective assessment only used the MPEG-2 4 Mbps, H.264 2 and 4 Mbps and H.265 1, 2 and 4 Mbps PVS of the video clip "FoxandBird." Thus $1 + 2 + 3 = 6$ impaired versions plus the original video clip were tested when the clip was in the *low*, *medium*, or *high bandwidth* set. The subset of PVSs shown in Table I results in a total of $(5\times13)+(15\times7) = 170$ stimuli to be used in the subjective assessment.

The source file for each video clip was 10 seconds long and was stored as an uncompressed YUV file in the UYVY 4:2:2 format in 1080p 29.97 fps. The source files were processed in 2013 and 2014 with an early commercial product that contained an implementation of each codec.[1] The manufacturer's default settings (excepting the requested bitrate) were used to encode each sequence. The compressed video sequences were then decoded using the same product and the resulting bitstream files were stored. The bitstream files decompressed by the MPEG2 and AVC codecs were stored as UYVY 4:2:2 files. The files decompressed by the HEVC codec were stored as YV12 4:2:0 files and then converted to the UYVY 4:2:2 format.

Video clips were played with constant timing so that up to four viewers could participate simultaneously. The clips were rated using the ACR method. Information pertinent to the test was displayed in video clips with a 50% gray background and white font. These video clips identified session or clip numbers and instructed viewers to record their votes (for example: "Session 2", "Clip 27", and "Please enter your vote for clip 27."). For brevity, we refer to the session number video clip as SNVC, the clip number video clip as CNVC and the vote video clip as VVC. Video clips were concatenated in this order: SNVC, CNVC, PVS, VVC, CNVC, PVS, VVC, ..., CNVC, PVS, VVC. The SNVC was displayed for 7 seconds, the CNVC for 2 seconds, the VVC for 7 seconds. This created a 9 second gray screen between each 10 second PVS. Each session was saved to an uncompressed UYVY 4:2:2 video sequence.

Three unique PVS viewing orders were randomly generated in an attempt to mitigate order effects. Each viewing order was constrained such that no source clip was repeated twice in a row. In order to keep the session lengths tolerable the three viewing orders were split into three sessions each. Sessions 1 and 2 had 57 PVS and Session 3 had 56 PVS. Each session lasted roughly 19 minutes including voting time.

C. Subjective Test Implementation

The stimulus video sequences of each viewing order were stored on a striped four solid state drive array in a video workstation in order to facilitate flawless playback.[2] The Machina software program was used to play the uncompressed video through an AJA® Kona® 3G video board. The two on-board SDI video output ports were used to feed two 24 inch Sony® LMD-2450W monitors and two 26 inch Marshall® V-R261-DLW professional-grade video monitors. One Sony monitor and one Marshall monitor received the video signal using an SDI input port and then passed the signal on to the next Sony and Marshall monitor respectively

[1]Our thanks and gratitude to Steve Glennon of CableLabs for his assistance in creating PVSs for our test.

[2]Certain commercial equipment, software, and services are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is necessarily the best available for this purpose.

Figure 2. A photo showing the four viewing stations in our viewing room.

using an SDI output port. Thus playback to all four monitors was controlled using the Machina software on the video workstation.

We conducted the subjective experiment in a quiet, lighting-controlled room. There was no natural light and lighting conditions were consistent throughout the experiment. Each of the four professional-grade monitors were placed on one corner of a rectangular table such that a person viewing a monitor on one corner would not be able to see the video output from any of the other monitors. Chairs facilitating an upright sitting posture were placed such that a viewer's eyes were approximately 3 screen heights (3H) from the monitors and tape was placed on the floor to ensure consistent viewing distance throughout the test. Although the chairs were placed to facilitate a 3H viewing distance (as per ITU-T Rec. P.913), viewers were able to reduce the viewing distance by leaning forward. A photograph of the viewing room can be seen in Figure 2.

Experiment participants entered their votes for each video stimulus using a mobile device. The amount of light emitted by the mobile device was suitable for the purpose of entering votes but was not enough to distract from the task of watching the video clips. The mobile device wirelessly connected to a separate workstation running the Web-Enabled Subjective Test software (WEST) [11] in scoring-only mode. Scoring-only mode in the WEST software allows multiple experiment participants to record votes for an external stimulus. In this mode, the experiment administrator configures a mobile device to connect to the workstation running the WEST software and specifies the user and session number for which it will be recording votes. Once the test begins, the software displays "Clip 1" on the screen of the mobile device while the video plays back on the professional monitor. When the video clip has finished playing, the screen on the mobile device displays a voting interface. The experiment participant enters a vote, pushes submit, and then "Clip 2" is displayed on the screen of the mobile device until the second video stimulus is finished playing. The voting process repeats until all video stimuli for the current session have finished playing. The overall process was repeated for all three viewing sessions.

The experiment administrator read instructions from a script to each group of viewers in order to provide a consistent experience. The script included a description the format of the test, information about the testing room and its exits, and instructions on how to properly use the interface displayed by the WEST software on the mobile devices. A practice session was conducted before the first session in order to familiarize the viewers with experiment procedures. The practice session consisted of three video stimuli. The first and third stimuli were a high-quality (H.265 at 16 Mbps) and low-quality (MPEG-2 at 4 Mbps) version of a scuba diving scene not included in this test. The second stimulus was the scene depicted in Figure 1g coded using H.264 at 4 Mbps. This configuration allowed demonstration of the quality levels to be expected during the test while minimizing viewer boredom.

D. Subjective Assessment

A total of 25 viewers participated in the experiment; 12 were male and 13 were female. There were 3 viewers in the 15–24 year old age group, 7 viewers in the 25–34 age group, 4 viewers in the 35–44 age group, 5 in the 45–54 age group, 4 in the 55–64 age group and 2 viewers were older than 65. One viewer was identified as an expert viewer and his results were included in the overall test results. Each viewer completed a color vision deficiency test and one male was found to have abnormal results. His votes were included in the overall test results. This should have resulted in $25 \times 170 = 4,250$ votes but one viewer did not properly submit votes for all videos resulting in 4246 total votes cast.

E. Objective Assessment

Objective video quality scores were calculated for all 170 PVSs using two metrics: PSNR and the NTIA Video Quality Metric (VQM) [5]. In both instances the full-reference calibration method was employed. VQM reports measurements on a scale from 0 to 1, 0 meaning a flawless reproduction of the original sequence and 1 meaning the lowest possible quality. Because PSNR reports measurements in decibels and decibels are not directly related to subjective quality, we map PSNR results onto the same scale as VQM using the equation in Section 6.5 of [5]. Then both the PSNR and VQM scores were linearly mapped from the $[0, 1]$ scale to the typical $[1, 5]$ ACR scale resulting in PSNR_{MOS} and VQM_{MOS} values. By examining these objective scores, we can understand each model's response to the three codecs.

IV. Results

A. Subjective Assessment Results

Before calculating the results of this test, we analyzed the subjective data to look for anomalies. Calculating the correlation of one subject's votes for each PVS to the mean of all other subject's votes for each PVS is one way to check

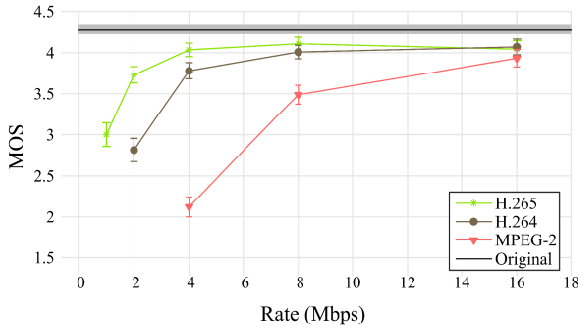


Figure 3. Quality level achieved by each codec using all available data and 95% confidence interval by bitrate. Shaded gray area indicates 95% confidence interval for overall original source MOS. Bitrates in megabits per second.

if a subject fundamentally understood the task at hand. Two viewers' correlation values were calculated to be $\rho = 0.16$ and $\rho = 0.14$ while correlation values for the rest of the viewers ranged from $0.49 \leq \rho \leq 0.81$. These viewers are clearly outliers and their results were therefore excluded from our test (as per Annex A.1 of ITU-T Rec. P.913), resulting in $(23 \times 170) - 4 = 3906$ votes used for analysis.

Analysis of the subjective test data generally supported the claim that H.265 can provide the same quality as H.264 at half the bitrate. Figure 3 shows the MOS for each codec at each bitrate tested averaged over all scenes using all available data. We can see that MPEG-2 doesn't come close to quality saturation until 16 Mbps while H.264 approaches saturation at 8 Mbps and H.265 approaches saturation at 4 Mbps. The MOS achieved by H.264 16 Mbps is 4.1 and the MOS achieved by H.265 8 Mbps is 4.1. Figure 3 shows that the confidence intervals overlap and therefore further testing is necessary to determine if the two MOS scores are statistically similar.

We performed the two-sample Student's *t*-test for each case represented by a Roman numeral shown in Tables II and III. For each case in Table II we employ the null hypothesis that the H.264 n Mbps votes and the H.265 $\frac{n}{2}$ Mbps votes for each scene come from the same distribution. For each case in Table III we employ the null hypothesis that the MPEG-2 m Mbps votes and the H.265 $\frac{m}{4}$ Mbps votes for each scene come from the same distribution. For example, the two-sample *t*-test comparing H.264 16 Mbps to H.265 8 Mbps is represented by case IV in Table II. During the subjective assessment votes for the *all impairments* set and the *high bandwidth* set processed with H.264 16 Mbps set were collected. Similarly, votes were collected for the *all impairments* set, the *medium bandwidth* set and the *high bandwidth* set all processed with H.265 8 Mbps. However, only the *all impairments* source set and the *high bandwidth* source set processed with H.265 8 Mbps were compared to the same sets processed with H.264 16 Mbps. In case IV, the null hypothesis is not rejected ($p = 0.67$). The Student's

Table II
H.264 AND H.265 COMPARISON TEST CASES AND ASSOCIATED P-VALUES AND MOS DIFFERENCES. ROMAN NUMERALS INDICATE TEST CASE INCLUSION, BULLETS INDICATE AVAILABLE DATA NOT INCLUDED IN A TEST CASE.

set	H.264				H.265			
	2	4	8	16	1	2	4	8
all	I	II	III	IV	I	II	III	IV
low	I	II			I	II	•	
med		II	III			II	III	•
high			III	IV			III	IV

	I	II	III	IV
p_2	0.07	0.49	0.08	0.67
p_p	0.003	0.31	0.01	0.51
Δ_{MOS}	−0.19	0.05	0.11	0.03

Table III
MPEG-2 AND H.265 COMPARISON TEST CASES AND ASSOCIATED P-VALUES AND MOS DIFFERENCES. ROMAN NUMERALS INDICATE TEST CASE INCLUSION, BULLETS INDICATE AVAILABLE DATA NOT INCLUDED IN A TEST CASE.

set	MPEG-2			H.265		
	4	8	16	1	2	4
all	I	II	III	I	II	III
low	I			I	•	•
med		II			II	•
high			III			III

	I	II	III
p_2	1.8×10^{-18}	0.40	0.03
p_p	8.5×10^{-32}	0.22	0.01
Δ_{MOS}	−0.88	−0.07	0.17

t-test indicates that H.264 16 Mbps is statistically equivalent to H.265 8 Mbps.

For all two-sample *t*-tests conducted in Table II, the null hypothesis was not rejected. The p-value for two-sample *t*-tests are indicated by p_2. Each two-sample *t*-test indicates that H.264 is statistically equivalent to H.265 at half the bit-rate. This supports the general rule-of-thumb that H.265 produces a quality equivalent to H.264 while using approximately one-half the bit-rate.

Looking at Table III, the null hypothesis was not rejected in case II but was rejected for cases I and III. Investigating case I, the MOS achieved by MPEG-2 is 2.12 and the MOS achieved by H.265 is 3.0. For case III the MOS achieved by MPEG-2 is 3.9 and the MOS achieved by H.265 is 3.8 using only the *all impairments* and *high bandwidth* sets. Thus we can say that H.265 at a quarter of the rate performed better than MPEG-2 in case I, as well as MPEG-2 in case II, and worse than MPEG-2 in case III. Though this result seems to contradict the data in Figure 3, it is important to remember that Figure 3 was created using all data available in the test. The results of the two-sample *t*-test use specific subsets to facilitate fair comparisons.

Using the sets of data as described in the two-sample *t*-test analysis we also performed paired *t*-tests for each case represented by a Roman numeral shown in Tables II and III. Instead of analyzing two separate distributions the paired

t-tests analyzed the distribution of the difference of votes matched by subject and source sequence. Votes for H.265 were always subtracted from votes for H.264 or MPEG-2. For each case in Tables II and III we employ the null hypothesis that the mean of difference of votes is equal to zero. The p-values for paired *t*-tests are indicated by p_p in Tables II and III. The value Δ_{MOS} is also given and was calculated by subtracting the MOS for each case for H.265 from the MOS for each case for H.264 or MPEG-2. Δ_{MOS} is equivalent to the mean of the differences calculated for use in the paired *t*-tests.

In Table II p_p indicates that for cases I and III the null hypothesis was rejected. For case I the mean of the difference (or equivalently, Δ_{MOS}) is -0.19 indicating that H.265 at 1 Mbps outperformed H.264 at 2 Mbps with 95% confidence. For case III the mean of the difference is 0.11 indicating that H.264 at 8 Mbps outperformed H.265 at 4 Mbps with 95% confidence.

In Table III p_p indicates that for cases I and III the null hypothesis was rejected. For case I the mean of the difference is -0.88 indicating that H.265 at 1 Mbps outperformed MPEG-2 at 4 Mbps with 95% confidence. For case III the mean of the difference is 0.17 indicating that MPEG-2 at 16 Mbps outperformed H.265 at 4 Mbps with 95% confidence.

Using paired *t*-tests allows us to leverage our experiment design to draw more specific conclusions. These results do not allow for a sweeping conclusion to be made and it is also difficult to extract an overall performance trend. However, the paired *t*-test for case III in Table II tells us that H.264 outperforms H.265 by 0.11 points on a ACR scale (or 2.75% of the scale). Even though these statistical tests have allowed us to resolve these performance differences, they are small and it would be difficult to adjust coding parameters to compensate.

We also analyzed the subjective assessment data on a per-scene basis and Figure 4 focuses on MOS for H.264 8 Mbps and H.265 4 Mbps across all scenes. H.264 8 Mbps was statistically better for scenes 'c' and 'm,' H.265 4 Mbps was statistically better for scene 'e.' The overall trend of estimated coding complexity is approximately borne out and is shown in Figure 4—the lowest scores are reported for the *all impairments* set, the highest scores are reported for the *low bandwidth* set, etc.

Figure 4 underscores the importance of diverse scene selection when designing a subjective test. The range of MOS spanning all scenes is roughly 1.5 points. A coding rate can deliver excellent quality for one type of content but only fair quality for another type. This is an important factor to consider when selecting bitrates for video transmission.

B. Objective Assessment Results

Figure 5 shows a scatter plot comparing VQM_{MOS} and MOS for each PVS. We cannot use the per-codec means to compare codec performance directly, because different

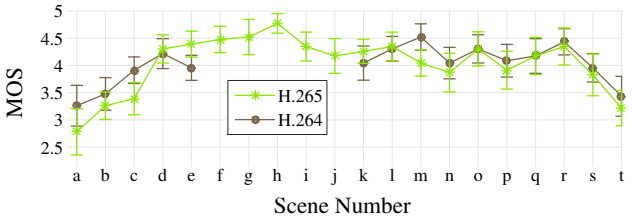


Figure 4. MOS for each scene as processed by H.265 4 Mbps and H.264 8 Mbps.

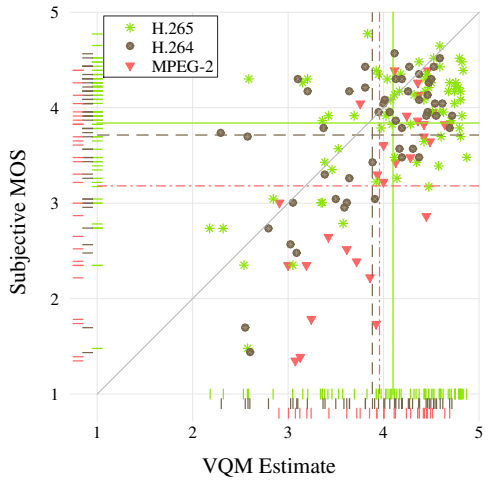


Figure 5. Scatter plot showing objective estimates for each PVS versus MOS for each PVS. Vertical lines represent means for each codec's objective estimate, horizontal lines represent means for each codec's MOS. Overall correlation is $\rho = 0.60$.

subsets of PVSs are associated with each codec. The overall correlation achieved is $\rho = 0.60$, but VQM most closely correlates with MPEG-2 ($\rho = 0.75$) compared to H.264 ($\rho = 0.63$) and H.265 ($\rho = 0.60$). This is not unexpected because VQM was trained on MPEG-2 but not H.264 or H.265.

Figure 6 shows a scatter plot comparing PSNR_{MOS} and MOS for each PVS. For presentation purposes, the PSNR data obtained in Section III-E was further scaled using $\overline{\text{PSNR}_{\text{MOS}}} = (\text{PSNR}_{\text{MOS}} \times 2) - 5$ as PSNR_{MOS} was contained in the interval $(3, 5)$. PSNR_{MOS} achieved a correlation of $\rho = 0.64$ to the subjective results but correlates most closely to H.265 ($\rho = 0.64$) compared to MPEG-2 ($\rho = 0.62$) and H.264 ($\rho = 0.60$). Note the consistent, low correlation numbers for each codec. Figure 6 indicates that PSNR is unreliable when comparing H.265 to the earlier codecs. This is consistent with the findings of Huynh-Thu and Ghanbari [12], who conclude that PSNR is only suited for comparisons within one video sequence and one codec, for example, optimizing encoding parameters for one clip.

Though these objective estimates are certainly not uncorrelated, they are also not a suitable replacement for subjective tests. Neither VQM nor PSNR should be used to compare the performance of H.265 with either H.264 or

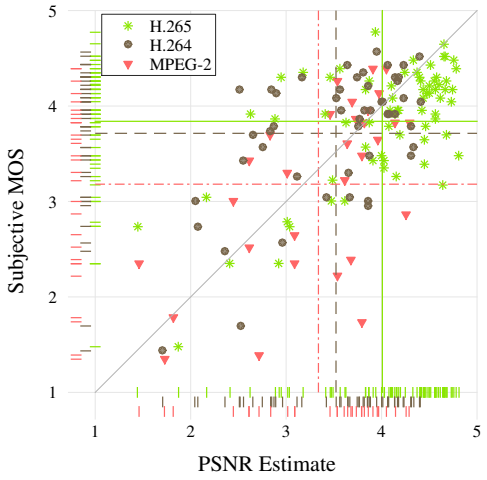


Figure 6. Scatter plot showing objective estimates for each PVS versus MOS for each PVS. Vertical lines represent means for each codec's objective estimate, horizontal lines represent means for each codec's MOS. Overall correlation is $\rho = 0.64$.

MPEG-2.

V. Conclusion

This analysis used a commercial implementation of MPEG-2, H.264, and H.265 to compare overall system quality based on 10 to 15 high definition (HD) scenes in the 1080p 29.97 fps format. Where H.264 was compared with H.265 at half the bitrate, H.265 performed as well as H.264 or better in three out of the four cases. In the case where H.264 performed better than H.265 the performance gap was only 2.75% of the ACR scale. This case also lies between two cases where the quality was equivalent. Where MPEG-2 was compared with H.265 at one-quarter bitrate, H.265 performed as well as MPEG-2 or better in two out of three cases. Notably, 1 Mbps H.265 ouperformed 4 Mbps MPEG-2 by 0.88 MOS points (or 22% of the ACR scale). In the case where MPEG-2 performed better than H.265 the performance gap was 4.25% of the ACR scale. However, we show that coding efficiency is significantly affected by content types and can cause an exception to the factor of two rule-of-thumb. That said, this independent analysis generally agrees with the efficiency savings reported by MPEG.

Neither objective metric, VQM nor PSNR, correlated well with our subjective results. Thus, the importance of conducting subjective tests is underscored.

The experiment described in this paper is part of a larger experiment in-which the ACR scale and a paired comparison method were both used to measure video quality. We will publish the results of this comparison in future papers. Afterward, the dataset will be made available at www.cdvl.org [10].

References

[1] *Advanced video coding for generic audiovisual services, ITU-T Recommendation H.264, 2014.*

[2] *Information Technology—Generic coding of moving pictures and associated audio information: Video, ITU-T Recommendation H.262, 2012.*

[3] *High efficiency video coding, ITU-T Recommendation H.265, 2015.*

[4] M. Pinson, S. Wolf, and G. Cermak, "HDTV Subjective Quality of H.264 vs. MPEG-2, With and Without Packet Loss," *IEEE Transactions on Broadcasting*, vol. 56, no. 1, pp. 86–91, March 2010.

[5] S. Wolf and M. Pinson, "Video quality measurement techniques," Institute for Telecommunication Sciences, Tech. Rep. TR-02-392, June 2002.

[6] J. Ohm, G. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—including high efficiency video coding (HEVC)," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1669–1684, December 2012.

[7] R. Weerakkody, M. Mrak, V. Baroncini, J.-R. Ohm, T. K. Tan, and G. Sullivan, "Verification testing of HEVC compression performance for UHD video," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, December 2014, pp. 1083–1087.

[8] R. Garcia and H. Kalva, "Subjective evaluation of HEVC and AVC/H.264 in mobile environments," *Consumer Electronics, IEEE Transactions on*, vol. 60, no. 1, pp. 116–123, February 2014.

[9] M. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2D and 3D subjective video quality tests," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 50, 2013. [Online]. Available: http://jivp.eurasipjournals.com/content/2013/1/50

[10] M. Pinson, "The consumer digital video library," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 172–174, July 2013.

[11] A. Catellier and L. Connors, "Web-enabled subjective test (WEST) research tools manual," Institute for Telecommunication Sciences, Tech. Rep. HB-14-501, January 2014.

[12] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, June 2008.