QoE Modelling for VP9 and H.265 Videos on Mobile Devices

Wei Song*, Yao Xiao*, Dian Tjondronegoro*, Antonio Liotta§
* Queensland University of Technology (QUT), Australia
§ Eindhoven University of Technology, The Netherland
{w1.song, dian}@qut.edu.au, pacifact@gmail.com, a.liotta@tue.nl

ABSTRACT

Current mobile devices and streaming video services support high definition (HD) video, increasing expectation for more contents. HD video streaming generally requires large bandwidth, exerting pressures on existing networks. New generation of video compression codecs, such as *VP9* and *H.265/HEVC*, are expected to be more effective for reducing bandwidth. Existing studies to measure the impact of its compression on users' perceived quality have not been focused on mobile devices. Here we propose new Quality of Experience (QoE) models that consider both subjective and objective assessments of mobile video quality. We introduce novel predictors, such as the correlations between video resolution and size of coding unit, and achieve a high goodness-of-fit to the collected subjective assessment data (adjusted R-square >83%). The performance analysis shows that H.265 can potentially achieve 44% to 59% bit rate saving compared to H.264/AVC, slightly better than VP9 at 33% to 53%, depending on video content and resolution.

Categories and Subject Descriptors

H.5.1 [Multimedia Information Services]: Evaluation/ methodology; **H.1.2 [User/Machine Systems]:** Human factor.

General Terms

Measurement, Performance, Human Factors and Verification.

Keywords

H.264/AVC, VP9, H.265/HEVC, video quality assessment, QoE modeling, mobile device.

1. INTRODUCTION

According to a recent report [1], global mobile video traffic has exceeded 55% of the total mobile data consumed in 2014. The majority of video contents are predicted to be mainly consumed on mobile devices such as tablets and smart phones on data network in the coming years [2]. This shift creates a sharp increase of demand for data bandwidth, pressuring the networks. To alleviate this problem, new video coding codecs are being developed, including VP9 by Google and H.265/HEVC (High Efficiency Video Coding) by the Joint Collaborative Team on Video Coding (JCT-VC) [3]. Accurate Quality of Experience (QoE) prediction models [4, 5] are crucial to estimate the potential benefits of using these new codecs to maintain or improve video quality on mobile devices, while using the same or less bandwidth compared to H.264/AVC.

MM '15, October 26-30, 2015, Brisbane, Australia
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2733373.2806256

The current research gaps are: 1) limited understanding on how much the new video codecs can increase the user's perceived quality on mobile settings; and 2) lack of QoE models for prediction of subjective quality based on objective measures. This paper aims to address both issues, which consequently will help to establish the potential benefits of VP9 and H.265 video codecs and inform future studies that use these codecs to achieve a scalable and user-centric HD video streaming [6-8]

We used subjective quality assessment method *Absolute Category Rating* (*ACR*), and objective quality assessments *SSIM* and *PSNR*, to benchmark H264, H.265 and VP9. We introduced novel predictors that define the characteristics of the video encoders, in addition to the existing video contents and encoding settings. These predictors were tested and selected to form two types of models using Linear Regression. One model uses objective quality measures as predictors, while the other does not. Both can explain over 83% of the total variation in subjective ACR scores. Our key contributions are:

- In-depth analysis on the subjective and objective assessment of VP9 and H.265 video quality on mobile devices, extending findings from existing studies, including the potential bitrate saving.
- Novel QoE models to predict users' perceived quality on mobile devices, across VP9, H.265 and H.264 codecs, as well as introduce some uniquely identified predictors.

2. RELATED WORK

The literature suggests that both H.265/HEVC [3] and VP9 video codecs are capable of saving approximately half of the required bandwidth while maintaining the same level of quality on conventional viewing devices such as TV, compared to the previous generation H.264/AVC [9] [10, 11]. However, the applicability of these findings on mobile devices is limited. For new codecs such as VP9, subjective quality assessment (sVQA) studies have been done mainly on conventional devices [12]. This limitation is due to the technical challenges of H.265 and VP9 videos playback on mobile devices because of the coding complexity. VP9 is a recent codec by Google and has only been adopted by few studies. Objective studies on VP9 bitrate saving performance [11, 13] have obtained inconsistent results, some of which are lower than the over 50% saving indicated in [11]. VP9 encoder was even shown to be inferior to the H.264 encoder with an average bitrate overhead of 8.4% at the same objective quality [13]. However, a subjective study [12] shows that VP9 performs the best for synthetic video content. Hence, further research is necessary to evaluate and verify the performance of VP9. QoE modelling for VP9 videos has not been done.

Compared to VP9, H.265 has attracted more attention and its performance has been investigated by both sVQA and oVQA assessments [12, 14, 15]. Current studies have established that H.265 can achieve high compression efficiency for Ultra High Definition (UHD) videos. However, only few studies have focused on mobile devices. One study [16] evaluated H.265 performance on a 4.5 inch mobile phone with video encoded at 360p and 200-400Kbps and suggested that H.265 has similar user experience to H.264 in low bandwidth settings. Yet this study cannot represent the current settings, as most of the new mobile devices can support much larger display resolution (e.g. 1920×1080pixels) and stream at higher rates (e.g., 4G, WiFi). Recent studies have modeled QoE for H.265 video content [17, 18] but not for mobile videos.

For objective video quality assessment (oVQA), the commonly used metrics are: peak signal-to-noise ratio (PSNR); Structural Similarity Index (SSIM); Multi-Scale SSIM index (MS-SSIM) Video Quality Matrix (VQM); and Motion-based Video Integrity Evaluation index (MOVIE). As PSNR only measures the signal distortions, many studies have found it to be significantly worse than all the other oVQA metrics for predicting subjective video quality. However, PSNR can still be a reliable metric if the impacts of video content and codec type are properly considered [19, 20]. SSIM can attain better prediction outcome because it makes use of the characteristic of human visual system (HVS), which is extremely sensitive to structural information of pictures, [21]. PSNR and SSIM are widely used for benchmark in various video quality studies. Using oVQA models on H.265 video quality assessment, study [14] found that all the existing models have underestimated the coding gain of H.265 compared to the increased scores from sVQA. This raises the need for future refinement of oVQA models.

For sVQA, the commonly used methods are set by ITU (International Telecommunication Union) [22, 23]. Single Stimulus Continuous Quality Evaluation (SSCQE) and Absolution Category Rating (ACR) ask viewers to rate one impaired video stream. Double stimulus continuous quality scale (DSCQS) and double stimulus comparison scale (DSCS) ask viewers to rate the quality or change in quality between two video streams (reference and impaired). Compared to DSCQS and DSCS, the results generated from ACR and SSCQE do not require the presence of reference video and are less resource and time consuming to carry out, but they are likely to be less accurate due to memory effect in test videos ordering. This drawback can be reduced by randomizing the video orders [24]. For mobile video assessment, double stimulus methods are not feasible for small-form factor display devices due to the lack of means to synchronize the screens from two mobile devices.

QoE models, as objective measurement of subjective video quality, are usually built upon the relationship of sVQA scores with some influencing factors/aspects, which are often considered as predictors. Generally, the predictors can be grouped into: Video signal distortion [21]; Encoding related parameters, such as spatial resolution, frame rate, bitrate and Quantization Parameter (QP) [18, 20]; Video content information, such as content type, and content characteristics of spatial and temporal complexity [18, 22]; Network parameters, such as packet loss rate, delay jitter and error rate [17, 25]; User interface information, such as screen resolution of mobile device or display resolution [20, 26]; and Codec information, such as codec type [26].

The abovementioned oVQA metrics (e.g., PSNR, SSIM, VQM, etc.) have been used in QoE models by many studies. These measurements are based on quality distortions of the received video relative to the original video, and how it can affect human visual system. However, these models do not incorporate knowledge about the video processing and delivery system, thereby more suitable for codec comparison and optimization [27]. Our study adopted the simplest models - PSNR and SSIM for comparing the performance of codecs, VP9, H.265 and H.264, and take further steps to improve their prediction accuracy.

Existing QoE models for H.265/HEVC videos have not leveraged much of other relevant information. For example, content-based video quality prediction (CVQP) model can be based on QP and content feature defined by motion and picture complexity index, which are extracted from the H.265 bit stream while encoding [18]. A further example is spatial-temporal characteristics of video content by using the weighted averages of coding unit quad-tree depth and prediction mode sequences [25]. Despite the benefits of incorporating these important factors of video content, the resulting models become codec dependent. A more generic method for defining video content characteristics is to calculate Temporal Information (TI) and Spatial Information (SI), which is recommended in ITU-T Rec. P.910 [22]. This method may not be profoundly accurate, but it is the most robust as it only associates with the original input video. Another study analyzed QoE for streamed H.265 videos from the impact of encoding bitrate, video content and transmission packet loss on subjective opinion [17]. It found that the combination of content type and packet loss has a significant impact on QoE. However, it has not provided a measurement of this impact.

Based on the literature, it is clear that many potential predictors still need to be verified for modeling the QoE of new generation codecs H.265 and VP9. In particular, in this paper we would identify the parameters to indicate codec types, and explore the potential correlations between video spatial resolution and codec-related information, which have not been discovered.

3. EXPERIMENT DESIGN

The study aims to construct a video quality prediction model for videos encoded by the latest codecs on mobile devices, using the data gathered from both subjective and objective video quality assessment methods. Figure 1 illustrates the study workflow.

Firstly, a set of master video sequences is selected from the IRCCyN 1080i HD video database [28] and prepared as reference sequences by scaling and de-interlacing. Secondly, compressing reference sequences into 5 different bitrates using 3 video encoders produce the distorted video sequences. Thirdly, commonly used sVQA method (ACR) and oVQA methods (PSNR and SSIM), are adopted to collect the video quality data, which subsequently is used to analyze the performance of the three codecs (H.265, VP9 and H.264). Finally, QoE models are constructed based on linear regression analysis, with the average ACR scores as the dependent variable and some encoding parameters and oVQA scores as the independent variables, i.e. the predictors. It should be noted that due to technical constraints (no available video player can directly playback H.265 smoothly), sVQA only assess H.264/AVC and VP9 encoders, while all three encoders will be assessed by oVQA. The subjective quality of H.265 are estimated based on the correlations between H.264, VP9 and H.265 in terms of objective quality and the correlations between H.264 and VP9 in terms of subjective quality. In our proposed QoE models, pixel format, frame rate and scanning mode are standardized, therefore will not be tested as predictors.

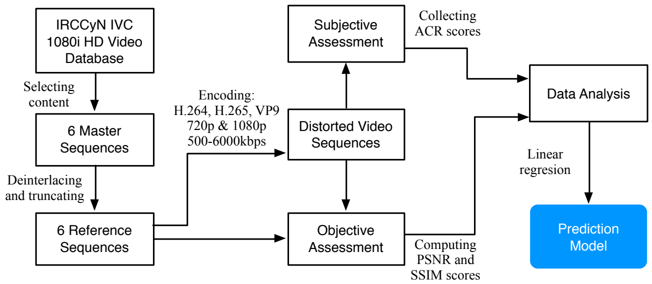


Figure 1. Overview of Video Quality Assessment Study Design

3.1 Video Source

We selected 6 videos as the master sequences to produce reference and distorted sequences. The selection was meant to cater the different scenes simulating real-life situations, which are described in Table 1. Table 1 also gives the average content complexity SI and TI value [22]. Refer to http://ivc.univ-nantes.fr/en/databases/1080i_Videos/ for more details about these video sources.

Table 1. Details of Selected Video Sequences

Seq	Source name	Description	Content Feature (SI&TI)
1	InsideMarathon	Large crowd runners moving from right to left without camera panning	Multiple fast moving objects with high spatial details in the still background (50,34)
2	Captain	Close-up of a captain looking through a monocular under heavy downpour	Close-up portrait with slow and minor zoom-in; low contrast (39,13)
3	Concert	A singer walking across the stage in spotlight with band members behind him.	Random panning and zooming; lowlight condition (43,16)
4	GroupDisorder	People running in different directions combined with camera movement	Fast random motion and panning camera (57,28)
5	Foot	Making a goal in a football match	Fast moving objects with wide angle camera panning and zooming (68,26)
6	DanceInTheWoods	People dancing on a trail in a forest	Random moving objects with camera zoom in; high contrast (57,18)

Master sequence 1 consists of extreme details and fast moving objects that will test the inter-frame compression capability of video encoders. Master sequence 2 contains close-up portrait, commonly seen in television drama series, talk shows, news broadcast and movies where protagonists' facial expressions are important. Master video sequence 3 was selected to demonstrate the recordings created under poorly and unprofessionally lit circumstances, such as candid shot, live music concert, and home videos created in uncontrolled environment. Such scene often contains dark background and gradient which will test the capability of encoders of dividing MBs. Master video sequence 5 was chosen as users often watch fast-paced sports games such as soccer, rugby and basketball where cameras are constantly panning with fast-moving players. Master video sequence 4 and 6 were selected to simulate video contents of high color contrast. Such contents are often seen in movies as they have been processed extensively by postproduction techniques such as colour correction to create the mood and atmosphere directors want to convey to audience. Despite such video contents are highly distorted due to editing and post-production, they are likely to be appealing to the eyes of audiences.

3.2 Reference and Distorted Video Sequences

Most online video contents are progressive scanned 4:2:0, therefore we de-interlaced and sub-sampled the 6 master sequences from YUV4:2:2 50i 1080p to YUV4:2:0 25p 1080p format by using FFmpeg [29] to generate the reference sequences. YUV4:2:0 sampling rate was adopted because existing literature suggests HVS is not as sensitive to pick color information as to pick up textual information [9], therefore, it is unlikely that end-users will be able to tell the difference in video quality between YUV4:2:0 and YUV4:2:2 sequences. Moreover, all online video content distributors, such *YouTube* and *DailyMotion,* currently deliver their videos in YUV4:2:0 sampling rate. YUV4:2:2 sampled sequences are neither easily available to end-users, nor sufficiently compact to be transmitted online.

The master sequences were de-interlaced in this study due to the decreasing number of video contents being delivered in interlaced scanning format to end-users, as modern flat panel displays on devices of all sizes are capable of displaying progressive scanning videos natively. Moreover, the latest generation of video encoders, such as the VP9 and H.265/HEVC have abandoned the support of field coding, as interlaced scanning is considered to be obsolete and no longer used for display devices and content distribution [3].

The 6 master sequences have at least ten seconds of duration. For standardization, we truncated them to be 10 seconds or 250 frames each to produce our reference sequences. The 6 reference sequences are then compressed into distorted sequences using the combinations of the following encoding settings:

- Bitrate (BR): 500kbps, 1Mbps, 2Mbps, 4Mbps, 6Mbps
- Resolution (RS): 720p and 1080p
- Coding Format (CF): H.264, H.265 and VP9

At the end, 180 distorted sequences in total were generated ($5BR \times 2RS \times 3CF \times 6 Sequences = 180\ combinations$).

While all three encoders used in this study are capable of sub-sampling, FFmpeg was used to pre-subsample (/down-scale) the reference sequences to 720p resolution before encoding, in order to prevent inconsistency that might be caused by potential difference of the resizing arithmetic in different video encoders. After the distorted or compressed video sequences are decoded by their respective video decoders, the distorted raw YUV file in 720p resolution will then be up-scaled by using FFmpeg for the purpose of comparison with the reference sequences in oVQA.

We adopted Variable Bitrate (VBR) setting as bitrate control mode for encoding as most of the online videos are encoded in VBR mode. Unlike Constant Bitrate (CBR), which uses same amount of data (bits) to interpret every frame of a given video regardless of the frame complexity, VBR is able to adjust the amount of data allocated to represent each frame based on the level of image complexity. Consequently, a video content encoded with VBR setting is likely to be superior to those encoded by CBR if the sizes of the video files are kept the same. Allowing bitrates to vary significantly in a given sequence will affect the overall size of the encoded video; therefore, we allowed only marginal fluctuation of bitrates and make sure the final file sizes of distorted sequences are corresponding to their bitrates. For instance, a 10-seconds long distorted sequence produced from its reference sequence with a bitrate setting of 500kbps (kilobit per second) is expected to have a file size of 610.3kB (kilobyte) approximately. We allowed a tolerance of ±5% on the expected final size of distorted video sequences.

3.3 Encoder Settings

Three video encoders x264 0.120, VP9 v1.2.0, HM14 were used to generate three coding formats H.264/AVC, VP9 and H.265/HEVC respectively. These encoders were the latest at the time of this study. The default encoder settings are different; therefore, we made adjustments to the encoding parameters to ensure a fair comparison as much as possible. The summarized configuration options are described in Table 2.

Table 2. Encoder configurations

Coding Format	**H.264/AVC**	**VP9**	**H.265/HEVC**
Encoder Version	FFmpeg 0.9.4 with x264 0.120	VP9 v1.2.0	HM14
Encoding value	Very slow	Max	–
Pass	1	1	1
Maximum coding block Size (pixels)	16 × 16	64 × 64	64 × 64
Bitrate Control Mode	VBR (defined target bitrate)	VBR (defined target bitrate)	VBR (defined target bitrate)
GOP Length (Intra Period)	320	320	320
GOP Size	Auto	Auto	8
Internal Bit Depth	8	8	8

The Group of Pictures (GOP) size of H.265 encoder is set to the default value of 8 as it is a compulsory option to be turned on. Since the number of inter frames encoded are impossible to be kept constant for all three encoders due to their design differences, we only specified the maximum intra-frame distance for all three encoders. Intra-frame is usually introduced at a scene change or a start of video sequence. As each of our video sequences contains only one scene, only one I-frame is encoded in each of the distorted sequence. VP9 encoder includes a golden-frame, which functions the same as intra-frame, but is not displayed by the decoder and cannot be toggled off in the encoder settings.

3.4 sVQA Protocol

3.4.1 Test Environment

Video sequences were presented on a Microsoft Surface Pro 1st Generation which comes with 10.6-inch screen of native resolution 1920 × 1080 pixels, Intel i5 CPU, 4GB memory, 64GB Solid-State Drive, and the up-to-date Windows 8.1 Pro OS. The only third party software installed onto the device were *Google Chrome* for running our web-based sVQA application and *Combined Community Codec Pack* (CCCP) for detecting playback statistics. The web-based sVQA application and the test video sequences were copied into the solid-state hard drive for smooth playback. The device is selected for its capability of decoding both VP9 and H.264/AVC sequences smoothly without skipping frames and visual latency. A wireless mouse was used for a fuss-free operation and familiarity, despite the availability of touch screen as user input interface. The mobile device was disconnected from the Internet during the assessment sessions.

3.4.2 Participants

The sVQA study was conducted at a controlled lab environment as recommended by ITU-T Rec P.910 [22]. All the participants were asked to sit in an enclosed space where there was no source of distraction or strong window light, but florescence. The viewing distance was about 80cm and the viewing angle was about 30 degree. During assessment, the participant was left alone to complete his/her tasks while we kept a distance for observation without disruption, to minimize the unwanted effects of being supervised, such as tense and indecisive assessment [30].

The criteria for participants are:

- Adult, 18-50 age range
- No experience with video quality assessment study
- No visual impairment or illness associated with visual capabilities, e.g. colour blindness, severe myopic, etc.

We have recruited a total of 30 participants aged between 20-35 from university students and staff. The number of participants meets the common requirement for sVQA, prescribed by Video Quality Experts Group [28, 31] and ITU-R Rec. BT 500-13 [23]. To confirm that all our recruited participants do not have visual-associated illness or impairment, we asked them to describe the colors of the objects shown in the training videos and to make simple quality comparison during a demo session. Participants who are unable to fulfill these tasks would be deemed as unsuitable to sit for our study, and are therefore removed from the study. No invalid participation was found in our study.

Each sVQA session was separated into two parts: training and the actual evaluation. To get the participants familiar with the operation of our sVQA application and hardware, our facilitator explained in details about how our application and hardware function to each of the participants before they got started with 2 sets of training sequences during demo sessions. The participants were allowed to try the demo session as many times as they desired until they were confident with the application interface and sVQA operations. After the demo session, participants could start the actual sVOA session at anytime. Participants were informed that they were allowed to terminate the session anytime they intended should they felt uncomfortable or tired. Their scores, under such circumstances, would not be recorded.

3.4.3 Voting method

We adopted ACR method [22] to gather sVQA scores, which uses a 9-point scale (1-bad to 9-excellent) to give a more precise rating than 5-point scale. No hidden reference video sequence was shown to the participants during the sVQA sessions. In each sVQA session, a participant viewed one distorted sequence at a time and rated the video quality using a 9-point slider after viewing each sequence. A limited voting time of 10 seconds was applied. No score was recorded if the subject failed to make decision during the given time. Participants in this study have gone through a short training (demo) session to minimize the chance of failing to vote.

The sVQA study only assessed video sequences encoded by H.264 and VP9 encoders without reference video presented. In total, 120 distorted sequences were used. To alleviate participants' fatigues, we divided the 120 sequences into six groups to be watched in six sessions, and a short 5-10 minutes break was allowed between the sessions. Each session corresponds to one master sequence and has 20 distorted sequences arranged in a random manner regardless of their bitrate, resolution and encoder. The presentation order of distorted sequences is identical to all participants in each session. Although the viewing order of the six sessions was not restricted, it was observed that most participants viewed the sessions in the order from 1 to 6.

We have designed an HTML5 web-based application for this study. The application can playback both VP9 and H.264 sequences through Google Chrome browser in full screen [32] and store the ACR scores given by participants into CSV files. The first view of the web application is a list of the six testing sessions and the demo session with hyperlinks. Clicking on each hyperlink will start to play the corresponding videos one after another. A rating screen is displayed immediately after each video sequence is finished. This screen includes: a sliding-bar labeled as 9-excellent, 7-good, 5-fair, 3-poor, and 1-bad to allow the participant voting; a 'Next' button beneath the sliding bar to confirm the vote; a 'countdown timer' on the top right-hand corner to prompt the participants the remaining time.

3.5 oVQA Computing and Analysis Tools

We used Video Quality Measurement Tool (VQMT) from the Multimedia Signal Processing Group (MMSPG) [33] to generate the oVQA scores. The tool generates both frame-by-frame and average PSNR or SSIM scores. In this study, each of the 180 distorted sequences was firstly decoded into YUV format and up-scaled to the same resolution 1080p by FFMpeg, and then paired up with its corresponding reference sequence to generate average PSNR and SSIM scores. These oVQA scores were stored into CSV files for statistical analysis.

To compare coding efficiency of different codecs or different coding settings, Bjøntegaard model [34] is adopted to calculate average bit rate difference in percentage (BD-BR) or average PSNR different in dB between two rate distortion (RD) curves. We also computed the average MOS (instead of PSNR) and bit rate differences between two sets of subjective values pertinent to two codecs, based on the Subjective Comparison of ENcoders based on fItted Curves (SCENIC) model [35].

IBM SSPS Version 22 and R were used to tabulate and analyze the retrieved subjective and objective scores. The analysis includes: 1) Identification of sVQA data outliers that affect the consistency of the dataset; 2) Comparison between sVQA and oVQA scores; and 3) Discovery of potential predictors and test regression results for prediction models.

4. RESULTS

4.1 sVQA Analysis Results

A total of 3597 ACR scores were recorded with 3 fail-to-vote records (120 test sequences × 30 participants – 3 fails = 3597). Outliers were identified through Boxplot for each combination of video content, bitrate, resolution and codec. As a result, 148 entries out of 3597 were flagged out as outliers hence removed from our ACR score list.

Figure 2 depicts the average ACR scores of distorted sequences across the 6 video contents at two resolutions, 720p and 1080p. The results show that VP9 encoder performs significantly better than the H.264/AVC encoder across all 5 bitrates. Both encoders demonstrated similar performance patterns when bitrate increases. The correlation between bitrate and ACR scores is logarithmic. The wide error bars shown in Figure 2 are caused by the difference of mean scores for 6 different video contents, which can be further explained by Figure 3. Table 3 lists the percentages of the average ACR score differences throughout all five bitrates by comparing VP9 to H.264/AVC encoders. The *t-test* results show the significance of ACR score differences.

Based on Figure 2 and the *t*-test results shown in Table 3, the VP9 encoder is statistically significantly superior to the H.264/AVC encoder based on subjective quality (p<0.025, except bitrate at 4000kbps). The mean ACR score received by the VP9 encoder is 80% higher than that of generated by the H.264/AVC encoder under 500kbps bitrate. However, the advantage of VP9 encoder diminishes gradually as the bitrate increases. Under the highest bitrate of 6000kbps, the VP9 encoder only received 6.8% higher ACR score on average than the H.264/AVC encoder.

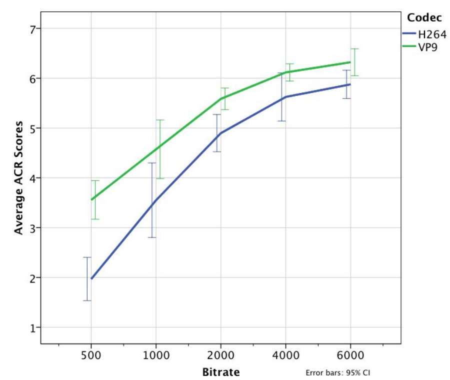


Figure 2. Average ACR scores of videos

Table 3. sVQA comparison between VP9 and H.264/AVC

Bitrate (kbps)	Mean Diff (%)	*t*-Test
500	+80	$t(11)=7.509$, $p<0.001$
1000	+27.8	$t(11)=3.248$, $p=0.008$
2000	+14.3	$t(11)=5.151$, $p<0.001$
4000	+8.9	$t(11)=1.909$, $p=0.083$
6000	+6.8	$t(11)=2.890$, $p=0.015$
Average	+19.1	

To analyze the impact of content and resolution on the subjective video quality, two separate diagrams are used for the six video contents and the two resolutions in Figure 3. Based on this figure, steep slopes are shown for content 1, 4 and 6, and relatively flat curves for content 2, 3 and 5. Based on the content description in Table 1, we conclude that higher bitrate contributing to better perceived quality is more for the videos with fast and large moving objects than for the videos with slow motion or small objects.

From the perspective of bit rate saving, the codec's performance (i.e. in terms of its subjective quality) is compared using the *SCENIC* model. Results in Table 4 show that the average bit rate reduction of VP9 relative to H.264 is 25.96% at 720p and 55.95% at 1080p, respectively. Moreover, H.264/AVC encoder appears to perform much better at 720p resolution than at 1080p resolution, gaining 62.65% of bit rate saving. VP9 encoder's performance is relatively consistent under the two resolutions and gains about 9.33% of more bit rate saving at 1080p than 720p. Their resolution related performance can be observed from Figure 4(a) and (b).

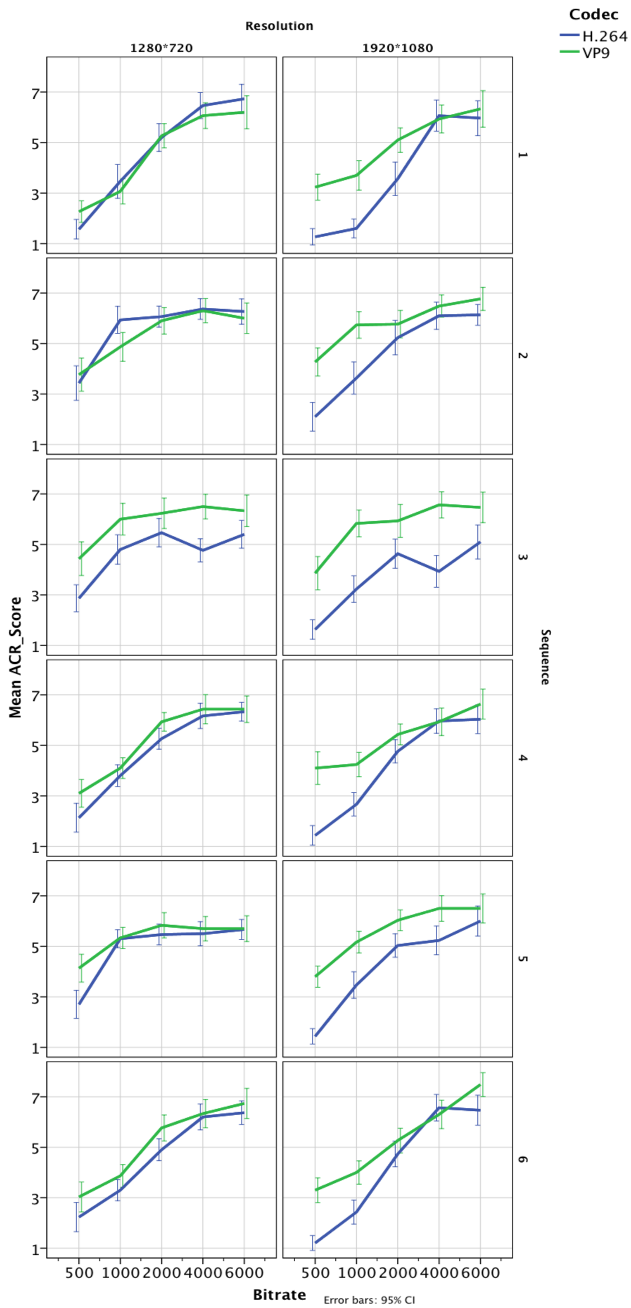


Figure 3. Average ACR scores for 6 contents and 2 resolutions

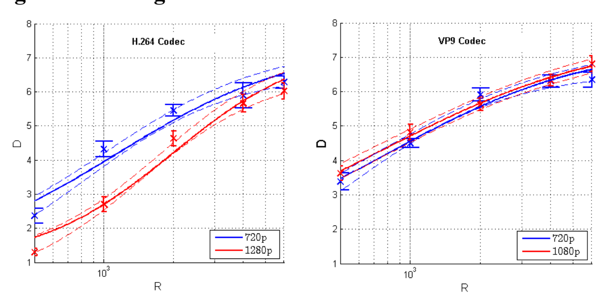


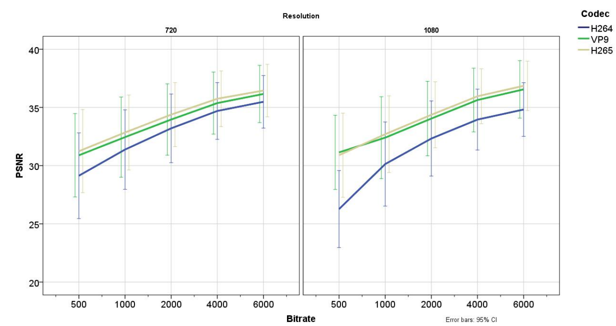
Figure 4. R-D curves of average subjective difference between two resolutions for (a) H.264/AVC and (b) VP9 codec

Table 4. Summary of bit rate saving with subjective comparisons between codecs and between resolutions

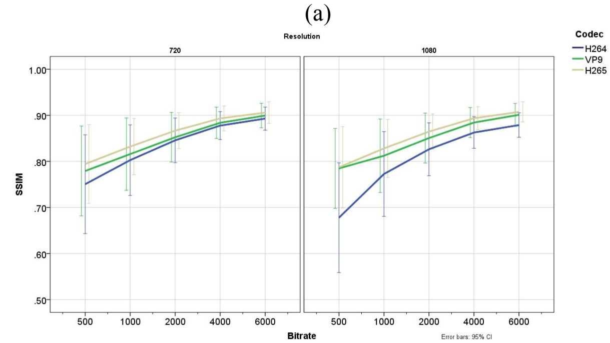
BD-MOS	H.264 vs. VP9		720p vs. 1080p	
	720p	1080p	H.264	VP9
Average	-25.96%	-55.95%	62.65%	-9.33%

4.2 oVQA Analysis Results

The PSNR and SSIM scores obtained are highly consistent and showing similar patterns, based on comparing Figure 5(a) with Figure 5(b). It can also be observed that: (1) The VP9 and H.265 have similar curves of oVQA vs bitrate for both 720p and 1080p resolutions, although the H.265 is overall marginally better than the VP9. (2) For both VP9 and H.265, the PSNR and SSIM scores under different resolutions are roughly equal (e.g., at 2000kbps, PSNR≈34.5dB, SSIM≈0.85). (3) The H.264 encoder performed poorly compared to the VP9 and H.265 encoders, especially at a relatively lower bitrate (e.g., 500kbps) and higher resolution. (4) The H.264 encoder at 720p performed significantly inferior to itself at 1080p. The observations of (2)-(4) are consistent with the sVQA results.



(a)



(b)

Figure 5. Averaged (a) PSNR and (b) SSIM for 6 contents

Further considering the observation (4), it is probably due to the fact that the H.264/AVC standard was designed in the early 2000s, before HD video contents became popular. It did not consider the spatial resolution in relation with the size of the supported MB and the data overhead caused by the number of MBs. The 16×16pixels MBs specified in the H.264/AVC standard will consume a big amount of MB overhead for a video with a large resolution; as a result, at a given encoding bitrate, more bits will be allocated in describing the MBs instead of storing actual frame information. For example, H.264 will generate a total number of 3600 (80 × 46 = 3600) MBs for a frame of 720p resolution, but 8160 (120 × 68 = 8160) MBs for the same video frame of 1080p, which is approximately 2.27 times of that of 720p. When the encoder allocates more binary bits to the overhead of MBs, it has to lower the amount of bits for the actual video information. Therefore, H.264/AVC encoder performed significantly worse at 1080p than 720p. Additionally, the negative effect caused by the MB overhead is more obvious under lower bitrate. On the other hand, to encode the video frame of 1080p, H.265 and VP9 will need only 510 (30×17=510) CUs/Superblock (64×64 pixels per block). This equates to approximately 15 times less bits for the overhead of a H.265/VP9 video stream than that of a H.264 video stream.

The PSNR and SSIM values are video content dependent, thus a wide error bars shown in Figure 5. Their trends of content dependency conform to those of ACR scores, that is, the PSNR or SIM value range becomes narrower with the lesser amount of motion and details in the videos. For example, The contents 1, 4 and 6 with a PSNR range of 25-37dB are distinguished from the contents 2, 3 and 5 with the range of 32-40dB across the bitrate from 500kbps - 6000kbps. Figure 6 compares the PSNR values of content 1 and 2 for different combinations of bitrates, resolutions and encoders.

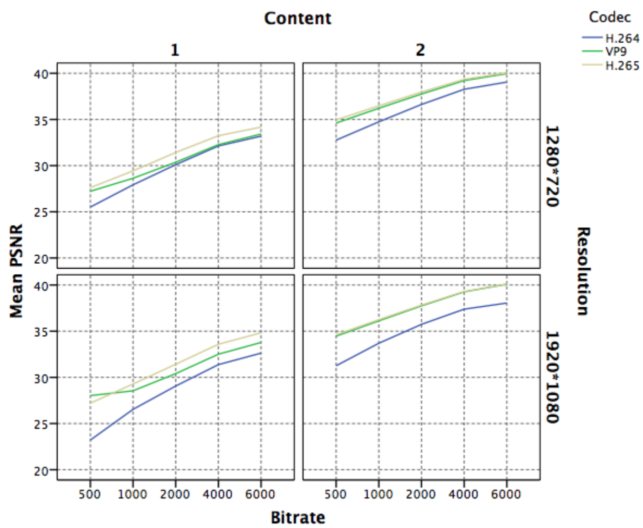


Figure 6. PSNR comparison between content 1 and 2

The Bjøntegaard metric was applied to calculate the bitrate saving achieved by the tested encoders based on the PSNR scores. The results, shown in Table 5, revealed that the average bitrate savings of the VP9 and H.265 encoders is 33% – 59% relative to the H.264 encoder. The bitrate saving varies with the video content type and the video resolutions. For example, video seq3 achieved the highest bitrate saving of 73.1% at 1080p resolution by using H.265 encoder; whereas seq1 achieved only 37.56% of bitrate saving. When the resolution increases from 720p to 1080p, the bit rate saving of H.265 or VP9 encoder versus H.264 encoder improves by about 15% or 19% consequently.

H.265 encoder is more efficient than VP9 encoder in terms of average bitrate saving: 5.85% at 1080p and 13.44% at 720p, especially for the video content that contains extreme motions and image details (e.g., seq1 and seq6). However, for less complex contents (e.g., seq2 and seq5), the advantage of the H.265 encoder over the VP9 encoder is very marginal, which is about 2-6% at 1080p. VP9 encoder outperformed the H.265 encoder by 2.3% under 1080p resolution for video that contains dark background, such as seq3. VP9 seems to have the performance advantage for high-resolution videos with low contrast images.

Table 5. Bitrate saving comparisons among three codecs

Parameter	Content	VP9 vs. H.264	H.265 vs. H.264	H.265 vs. VP9
720p @25fps	Seq1	13.94%	37.56%	28.91%
	Seq2	40.47%	46.16%	9.74%
	Seq3	57.04%	63.81%	0.33%
	Seq4	32.28%	38.36%	9.33%
	Seq5	31.84%	37.57%	9.23%
	Seq6	27.07%	43.12%	23.08%
	Ave	**33.77%**	**44.43%**	**13.44%**
1080p @25fps	Seq1	35.46%	52.83%	25.85%
	Seq2	57.83%	59.79%	3.93%
	Seq3	70.38%	73.13%	-2.32%
	Seq4	53.49%	56.38%	5.90%
	Seq5	54.45%	56.03%	1.79%
	Seq6	46.25%	55.88%	18.06%
	Ave	**52.98%**	**59.01%**	**5.85%**

5. QUALITY PREDICTION MODELS

Using the collected data from the sVQA and oVQA, we have constructed models to predict the user-perceived quality for the VP9 and H.265 encoders. The simplest mathematic model is a linear function, which is fast to compute and easy to use in real-world applications. Thus, we assumed the user-perceived quality, which values are the averaged ACR scores from 30 subjects (S_{ACR}), can be measured by a set of variables in a linear function. Linear regression analysis was applied in IBM SPSS tool for predictor testing and to generate the models. In linear regression, R^2 coefficient of determination is a common statistical measure of how well the regression line approximates the real data points. An R^2 of 1 indicates perfect goodness-of-fit. The pitfall of R^2 is that its value will only increase as more predictors are added to the regression model. To overcome this issue, adjusted R^2 should be used with models with more than one predictor variables. Another typical measure is standard error of the regression, also called the root mean square error RMSE. It indicates how close the observed data points are to the model's predicted values, of which lower value indicates a better fit. We also performed 3-fold cross validation for each model using R with 'DAAG' package 'cv.lm' function (http://www.stats.uwo.ca/DAAG). The prediction error is reported.

5.1 Potential Predictors

According to the logarithmic relationship between the ACR scores and bitrate, the ACR score could be represented as $S_{ACR} = a + b \cdot \log(bitrate)$, whereby the coefficient a determines the pass-through point of the curve $(1, a)$, while b determines the growth rate of the curve. Both coefficients may be influenced by encoder related characteristics and the video content and its encoding parameters, which have identified through the analysis in Section 4. To represent ACR scores as a linear model, the following potential predictors are considered based on their influences on video quality:

- *LBR*: the logarithmic bitrate with base 10. It converts the bitrate into a variable that is linearly related to ACR scores.
- *SR*: the spatial resolution of a given video divided by the resolution of 1280×720 pixels. (i.e., SR=1 for a 720p video).
- *LMB*: the largest size of macro block/coding unit that can be possibly allowed by a video encoder, divided by 16×16. (e.g., LMB equals to 16 for VP9, 1 for H.264).

- *RB*: the ratio of the SR to the LMB. It denotes the interactive influence between a video frame resolution and an encoder's coding characteristic.
- *TI*: the average value of temporal information over all frames of a video.
- *SI*: the average value of spatial information over all frames of a video.
- $LBR \times RB$ and $LBR \times TI$ are also considered to represent the impacts of the encoder and the video content characteristics on the curve of bitrate versus ACR scores.
- *PSNR*: the objective PSNR value divided by 30.
- *SSIM: the objective SSIM value.*

Then, ANOVA analysis is used to determine the significance of these potential predictors. The results in Table 6 show that SR, LMB and SI should be excluded because they did not significantly contribute to the model (p>.05).

Table 6. ANOVA results for the predictors

	Mean_sq	F	P
LBR	178.7	703.11	< 2e-16 ***
SR	0.3	1.11	0.29408
LMB	0.4	1.43	0.23480
RB	29.2	114.76	< 2e-16 ***
SI	0.7	2.80	0.09698
TI	4.8	18.73	3.4e-05 ***
LBRxRB	7.3	28.75	4.7e-07 ***
LBRxTI	3.4	13.57	0.00036 ***
PSNR	6.2	24.44	2.8e-06 ***
SSIM	4.0	15.86	0.00012 ***

*** Sig < .001

5.2 User-perceived Video Quality Modeling

When establishing the model for measuring the user-perceived video quality on small-form factor screen, we considered two situations: whether or not objective video quality measures (i.e., PSNR and SSIM) are available. We ran step-wise regression to build the final model. To ensure the model is robust, variables that have less than 1.5% of contribution to the model's R^2 change would be ignored.

For the situation of without oVQA parameters, after running step-wise regression, $LBR \times TI$, which increases the R^2 by 1.3%, is not included. Eventually, the model is generated as (1):

$$S'_{ACR} = 2.496LBR + 0.714RB \cdot LBR - 2.9RB - 0.028TI - 2.248 \qquad (1)$$

For model (1), R = 0.916 indicates a close correlation between the predicted ACR scores (S'_{ACR}) and the subjective ACR scores. Adjusted R^2 that measures the goodness-of-fit for the model is 0.834, which means 83.4% of the variations in ACR scores are captured by model (1). RMSE is also low at about 0.6. Cross-validation residual sums of squares is 0.377, which is the measure of the prediction error averaged across all three folds.

When taking the oVQA scores into consideration, TI is no longer needed as a predictor (p>.05), which is possibly because its influence on subjective quality is reflected by the objective measures. The model using PSNR as a predictor is presented in (2), and the model using SSIM is in (3).

$$S'_{ACR} = 1.925LBR - 0.567RB \cdot LBR + 3.573PSNR - 2.288RB - 5.059 \qquad (2)$$

$$S'_{ACR} = 1.786LBR + 0.502RB \cdot LBR + 6.691SSIM - 2.072RB - 6.244 \qquad (3)$$

The prediction performance of Model (2) is indicated by R=.930, Adjusted R^2=.861, RMSE=.555. 3 fold cross-validation shows its prediction error is 0.306. Model (3) performs even better, whereby R=.943, Adjusted R^2=.885, RMSE =.532, and 3-fold cross-validation prediction error is 0.248. Figure 7 demonstrates each fold's predicted values against the actual ACR variable based on model (3).

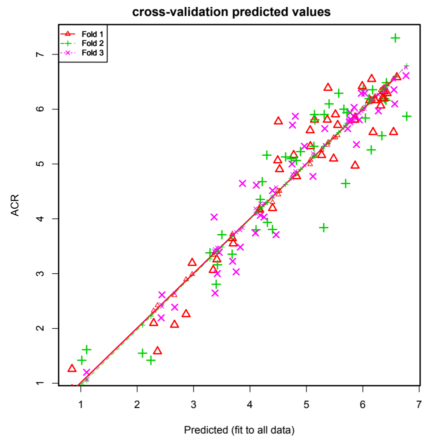


Figure 7. Scatter-plot of the predicted values by Model (3) versus subjective ACR scores

In all the three models, neither the video resolution (SR) nor the size of codec's largest macro block/coding unit (LMB) is significant for the user-perceived quality prediction, but their interaction (RB) shows as a significant predictor. The RB will increase as the video resolution increases while decrease as the size of LMB increases. Therefore, it is suggested to take this variable into consideration when developing subjective video quality prediction models across different codecs.

Although models (1) (2) and (3) are built with the ACR scores for VP9 and H.264 encoders, the format is applicable for other video codecs with similar block-based coding structure, because the predictor related to codec is generalized by the ratio of video frame resolution to the maximum size of the block. Only the coefficients of the model may be affected by the difference of codecs. Next, we will see this from the model for H.265 codec.

5.3 Subjective Video Quality Prediction for H.265

According to the oVQA result analysis, the objective performance of H.265 codec is similar to that of VP9, except it is marginally more efficient than VP9. Therefore, the subjective quality of H.265 videos is possibly predicted by using the correlation of the oVQA scores between the two codecs.

Based on the SSIM data, the SSIM values of videos encoded by H.265 can be represented by the values of VP9 in a close linear function (4), for which R^2 =.966, RMSE =0.01. The relationship is shown in Figure 8.

$$SSIM_{H265} = SSIM_{VP9} \times 0.885 + 0.11 \qquad (4)$$

By plugging (4) into (3), we can get the model (5), in which $0.11 \times LBR \approx 0.33$. Model (5) can be used to estimate the user-perceived quality of H.265 videos.

$$S'_{ACR} = 1.786LBR + 0.502RB \cdot LBR + 5.921SSIM - 2.072RB - 6.134 \qquad (5)$$

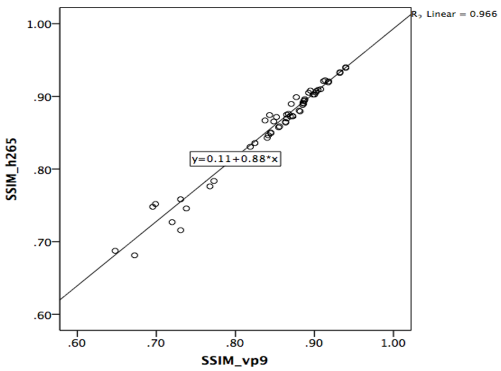


Figure 8. Correlations between the SSIM values from VP9 and H.265 video coding

6. DISCUSSION

In the sVQA and oVQA studies, we have examined the performance of VP9 and H.265/HEVC with the H.264/AVC as the baseline. Our quality assessment results are much different from the findings in [12], [13] and [16]. The detailed comparisons are show in Table 7. In [12, 13], H.265 shows dominant bit rate saving in comparison to VP9, which is about 35-43% in terms of objective PSNR and above 49% in terms of subjective score. Both studies state that VP9 is even worse than H.264/AVC with an average 5.1% of subjective loss or 8.4% of objective bit rate overhead. However, our study shows H.265 is only marginally superior to VP9, reaching an average of about 13% more bit rate saving at 720p resolution and downing to 6% when the resolution becomes 1080p. We concluded the same with [12, 13] in that H.265 is much better than H.264 with average 40-60% of bit rate saving according to Bjøntegaard Delta PSNR (BD-PSNR). This is different from [16], whereby no significant quality difference was perceived between the sequences compressed by H.265 and by H.264. This could be caused by the low bitrate 200-400Kbps as well as the small video resolution of 640×360 pixels used in [16].

We found that the advantage of VP9 and H.265/HEVC over H.264/AVC is dependent on both video content and video resolution, whereby videos with less motion and image complexity and bigger resolution will benefit more in bit rate saving than their counterparts. Content impact found in [12] shows HEVC is significantly better than the other codes in perceived quality of natural videos, while VP9 is the most efficient on synthetic video content. Our study used all natural videos; thus the advantage of VP9 on synthetic content is not captured. Instead, we discovered that VP9 performed the best on the content contains mostly dark or low contrast frames (e.g., seq3). The content dependency identified in [11] suggests that the number of points of visual attentions affects user experience. As the points of attention are related to the content complexity (e.g., SI and TI), this finding should be consistent with our result.

Interestingly, based on sVQA, the quality of H.264 video at a smaller resolution (e.g., 720p) is much better than it at a bigger resolution (e.g., 1080p), gaining an average 63% of bit saving. Therefore, if the H.264/AVC encoder has to be used for video compression, it is recommended to encode videos to 720p instead of full HD 1080p format at a given bitrate.

For QoE modeling, we considered two situations: without and with oVQA measurement as a predictor. The first situation is applicable for any cases; while the second can be regarded as further refinement of an existing oVQA metric (PSNR or SSIM) for QoE prediction. We determined the model predictors based on their significant impacts on subjective video quality. Except the bitrate related variable – logarithmic bitrate LBR, a special predictor RB was firstly identified as the ratio of a video resolution to the largest coding unit size supported by a video codec. This RB variable is used in both types of models, so is its interaction with LBR (RB×LBR). The QoE models involving such predictors can reflect codecs' performance variation with the changes of bitrate and video resolution. Thus, they are useful for predicting subjective quality of videos encoded by the codecs having block-based coding structure.

Table 7. Results comparisons with other studies

Items	[7]	[8]	[11]	Our Study
Study settings	4K UHD 1.5~20Mkbps 56" TV	832x480 ~ 2560x1600 QP: 22~37	640x360 200~400kbps 4.3'' mobile	720p, 1280p 500~6Mkbps 10.6" tablet
Method	Subjective Objective	– Objective	Subjective –	Subjective Objective
H.265/HEVC vs H.264	52.6% 39.6%	– 43.3%	Similar –	– 44%~60%
VP9 vs H.264	-5.1% 1.59%	– -8.4%	– –	26%~56% 34%~53%
H.265/HEVC vs VP9	49.4% 35.6%	– 39.3%	– –	– 6%~13%
Content dependency	H.265 is best for natural video; VP9 for synthetic video	–	Depend on points of attention	More bit rate saving on less motion and image complexity content
Resolution dependency	–	–	–	More bit rate saving on a bigger resolution

7. CONCLUSION AND FUTURE WORK

In this paper, we propose a set of novel subjective video quality prediction models based on the data from subjective and objective video quality assessment studies, which are applicable for mobile tablet and across the current and next generations of video codecs, i.e., H.264, H.265 and VP9. These models are presented as simple linear functions with only three predictor variables for each, but can explain more than 83% of variations of our collected subjective quality scores and have low standard errors (RMSE<.6).

To collect subjective and objective quality assessment scores, ACR method and PSNR and SSIM metrics were applied respectively on video sequences encoded with 500 – 6000kbps, 720p and 1080p, and VP9, H.265 and H.264 codecs. With the sVQA and oVQA results, we have evaluated the performance of these codecs. The results show that both VP9 and H.265 significantly outperform H.264 with average 26-56% of bit rate saving in terms of sVQA, and 33-59% of bit rate saving in terms of oVQA, whereby more bit rate saving is achieved at a higher video resolution. H.265 performs slightly better than VP9, with average 13% of bit rate saving at 720p resolution reduced to 6% at 1080p.

The limitation of this work is that the video ordering influence could not be avoided due to the nature of ACR study design. This can be improved by randomization of the video display orders across participants. As our subjective study was conducted with the assumption of perfect network transmission, our QoE models are not suitable for quality prediction under poor networks such as high packet loss. However, the SSIM or PSNR can still capture certain level of quality loss caused by network transmission.

In future work, we will carry out further subjective assessment study to evaluate the model for H.265 video quality prediction (eq. (5)) and take the frame rate parameter into consideration, which were not studied in this paper due to technical constraints (decoder availability). We will also make further comparisons with other QoE prediction models.

REFERENCE

1. Cisco. *Global mobile data traffic forecast update, 2014-2019*. Cisco white paper 2015. Available from: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html.
2. Ooyala, *Q1 2013 Video Index - TV is no longer a single screen in your living room.* 2013.
3. Sullivan, G.J., et al., 2012. *Overview of the high efficiency video coding (HEVC) standard.* IEEE Trans. Circuits Syst. Video Technol., 22,12, 1649-1668.
4. Brunnström, K., et al., 2013. *Qualinet White Paper on Definitions of Quality of Experience. Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeti*.
5. Menkovski, V., Exarchakos, G. and Liotta, A. 2011. *The value of relative quality in video delivery.* Journal of Mobile Multimedia, **7**(3): p. 151-162.
6. Menkovski, V. and Liotta, A. 2013. *Intelligent control for adaptive video streaming*. In *Proceedings of Int. Conference on Consumer Electronics (ICCE2013), IEEE*. 2013.
7. Agboma, F. and Liotta, A. 2006. *User centric assessment of mobile contents delivery*, In *Proceedings of the 4th Int. Conf. Advances in Mobile Computing and Multimedia* (Yogyakarta, Indonesia). IEEE, 121-130.
8. Song, W., Tjondronegoro, D. and Himawan, I. 2014. *Acceptability-based QoE Management for User-centric Mobile Video Delivery: A Field Study Evaluation*, In *Proceedings of the ACM MM*, ACM. 267-276.
9. Sullivan, G.J., Topiwala, P.N. and Luthra, A. 2004. *The H. 264/AVC advanced video coding standard: Overview and introduction to the fidelity range extensions*. In *the SPIE 49th Annual Meeting Optical Science and Technology*.
10. Ohm, J.-R., et al., 2012. *Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC).* IEEE Trans. Circuits Syst. Video Technol. 22,12, 1669-1684.
11. Bankoski, J., et al. 2013. *Towards a next generation open-source video codec*. In *IS&T/SPIE Electronic Imaging*. SPIE
12. Řeřábek, M. and Ebrahimi, T. 2014. *Comparison of compression efficiency between HEVC/H. 265 and VP9 based on subjective assessments*, In *Proceedings of SPIE Optical Engineering + Applications*, SPIE, San Diego, California, USA.
13. Grois, D., et al., 2013. *Performance Comparison of H. 265/MPEG-HEVC, VP9, and H. 264/MPEG-AVC Encoders*, In *Proceedings of the 30th PSC* (Dec 8-11 2013). IEEE .
14. Zeng, K., et al., 2013. *From H.264 to HEVC: Coding gain predicted by objective video quality assessment models*, In *Proceedings of the 7th VPQM* (Scottsdale, Arizona).
15. Sung-Ho, B., et al. 2013. *Assessments of Subjective Video Quality on HEVC-Encoded 4K-UHD Video for Beyond-HDTV Broadcasting Services.* IEEE Trans. Broadcast. 59, 2, 209-222.
16. Garcia, R. and Kalva, H. 2014. *Subjective evaluation of HEVC and AVC/H.264 in mobile environments.*, IEEE Trans. Consumer Electronics. 60, 1, 116-123.
17. Nightingale, J., et al. 2013. *Modeling QoE for streamed H. 265/HEVC content under adverse network conditions*, In *Proceedings of the 5th Int. Conf. on Wireless, Mobile and Multimedia Networks* (Beijing, China). IET. 249-253.
18. Anegekuh, L., Sun, L. and Ifeachor, E. 2013. *Encoding and video content based HEVC video quality prediction.* Multimedia Tools and Applications. 74:1795. Springer.
19. Huynh-Thu, Q. and Ghanbari, M. 2008. *Scope of validity of PSNR in image/video quality assessment.* Electronics letters, 44, 13, 800-801.
20. Song, W. and Tjondronegoro, D. 2014. *Acceptability-based QoE Models for Mobile Video.* IEEE Trans. Multimedia. 16, 3, 738 - 750.
21. Hore, A. and Ziou, D. 2010. *Image Quality Metrics: PSNR vs. SSIM*. In *Proceedings of ICPR 2010*.
22. ITU-T. 2008. *P. 910 Subjective Video Quality Assessment Methods For Multimedia Applications.*
23. ITU-R. 2012. Rec. *BT. 500-13: Methodology For The Subjective Assessment Of The Quality Of Television Pictures.*
24. Pinson, M. and Wolf, S. 2003. *Comparing subjective video quality testing methodologies*, In *Proceedings of SPIE Video Communications and Image Processing Conference*, SPIE. 573-582.
25. Nightingale, J., et al. 2014. *Deriving video content type from HEVC bitstream semantics*. In *Proceedings of SPIE Photonics Europe*.
26. ITU-T. 2007. *Opinion model for video-telephony applications* in *Quality of service and performance – Generic and user-related aspects*.
27. Takahashi, A., Hands, D. and Barriac, V. 2008. *Standardization activities in the ITU for a QoE assessment of IPTV.* IEEE Communications Magazine. 46, 2, 78-84.
28. Péchard, S., Pépion, R. and P. Le Callet. 2008. *Suitable methodology in subjective video quality assessment: a resolution dependent paradigm*. In *Proceedings of the 3rd IMQA*.
29. Bellard, F. and Niedermayer, M. 2012.*FFmpeg.*
30. Song, W., Tjondronegoro, D. and Docherty, M. 2010. *Exploration And Optimization Of User Experience In Viewing Videos On A Mobile Phone.* Int. J. Softw. Eng. Know. Eng. 20, 8, 1045-1075.
31. Hands, D. and Brunnstrom, K. 2007. *Multimedia Group Test Plan Draft Version 1.19*, Video Quality Experts Group.
32. Protalinski, E. 2013.*Google adds its free and open-source VP9 video codec to latest Chrome build*.
33. Hanhart, P. *VQMT: Video Quality Measurement Tool*. 2013 Available from: http://mmspg.epfl.ch/vqmt.
34. Bjontegaard, G. 2008. *Improvements of the BD-PSNR model.* ITU-T SG16 Q.6.
35. Hanhart, P. and Ebrahimi, T. 2014. *Calculation of average coding efficiency based on subjective quality scores.* J Vis Commun Image Represent. 25, 3, 555-564.