IEICE TRANS. COMMUN., VOL.E102–B, NO.12 DECEMBER 2019

PAPER

Transferring Adaptive Bit Rate Streaming Quality Models from H.264/HD to H.265/4K-UHD

Pierre LEBRETON[†a)], *Nonmember and* **Kazuhisa YAMAGISHI**[†b)], *Member*

SUMMARY In this paper the quality of adaptive bit rate video streaming is investigated and two state-of-the-art models, i.e., the NTT audiovisual quality-estimation and ITU-T P.1203 models, are considered. This paper shows how these models can be applied to new conditions, e.g., 4K ultra high definition (4K-UHD) videos encoded using H.265, considering that they were originally designed and trained for HD videos encoded with H.264. Six subjective evaluations involving up to 192 participants and a large variety of test conditions, e.g., durations from 10 sec to 3 min, coding-quality variation, and stalling events, were conducted on both TV and mobile devices. Using the subjective data, this paper addresses how models and coefficients can be transferred to new conditions. A comparison between state-of-the-art models is conducted, showing the performance of transferred and retrained models. It is found that other video-quality estimation models, such as VMAF, can be used as input of the NTT and ITU-T P.1203 long-term pooling modules, allowing these other video-quality-estimation models to support the specificities of adaptive bit-rate-streaming scenarios. Finally, all retrained coefficients are detailed in this paper allowing future work to directly reuse the results of this study.

key words: *quality of experience, adaptive bit rate streaming, audiovisual-quality-estimation models, subjective, monitoring, stalling*

1. Introduction

The use of adaptive bit-rate video streaming has greatly increased over the past years, and many users watch videos daily on multiple devices (e.g., TVs, smartphones, and tablets) across multiple services (e.g., YouTube, Netflix, and Twitch). However, due to variation in network conditions, perceived quality may degrade due to coding artifacts or stalling, which can result in users dropping services. To determine when the service is degraded, user quality of experience (QoE) [1] needs to be measured. Much effort has been put into connecting QoE to technical parameters such as bit rate, frame rate, resolution, and stalling events [2]. This has resulted in the successful development of audiovisual-quality-estimation models enabling the estimation of the QoE of adaptive bit-rate streaming [3]–[9]. These models are used in real-life services and measure in real time the experience of users across the network. This provides valuable information to service providers about user experience and makes it possible to dynamically balance network usage and quickly react to failures.

Considering the real-time constraint and the fact these models need to run on devices with very low processing abilities (e.g., routers, set-top boxes, smart TVs, smartphone apps), they need, in addition to being accurate, to have very low complexity. This challenging task is solved by defining specific scenarios: video encoded using H.264, using a resolution up to high definition (HD), and a frame rates up to 30 frames per second (in the case of the model described in ITU-T Recommendation P.1203 [5]). However, with the increasing number of 4K-ultra HD [10] (4K-UHD) videos and the introduction of new coding schemes, such as H.265, there is a need to extend the support to these new use conditions.

Therefore, this paper focuses on studying the consequences of increasing the scope of two models: the standardized ITU-T P.1203 model [5] and NTT audiovisual-estimation model by Yamagishi et al. [4], [11] to support a new codec (H.265/HEVC [12]), higher resolution (up to 4K-UHD), and higher frame rate (up 60 frames per second (fps)).

The contributions of this work are twofold: first, this study identifies which parts of the NTT model [4], [11] and ITU-T P.1203 models [5] need to be retrained, identifies new sets of coefficients, and investigates how previous models (including their coefficients) can be transferred to new conditions. Second, this work addresses the issue of temporal aggregation of per-second audiovisual quality estimation into an overall quality score. This analysis is done by using the long-term pooling modules from these two models together with other video-quality-estimation models (e.g., VMAF [13], BRISQUE [14], and NIQUE [15]). This analysis has two main outcomes. It first allows the studying of the applicability of long-term modules to video quality-estimation models other than those for which they were designed. This is important because adaptive bit rate streaming-video quality estimation is still recent, and only a few models can support large quality variation and stalling events. The second outcome is to show how previous models designed for 10-sec sequences can be applied to these new scenarios to enable easy reuse of past efforts on video-quality modeling and increase the number of available models that support such conditions.

To achieve this, six intensive subjective evaluations under various test conditions (video duration, coding condition, devices, etc.) are conducted.

This paper is organized as follows. Section 2 provides information on previous work addressing the QoE of adaptive bit-rate video streaming. Section 3 gives an overview of the models addressed in this study. Section 4 discusses the

Manuscript received February 18, 2019.
Manuscript revised May 16, 2019.
Manuscript publicized June 25, 2019.

[†]The authors are with NTT Network Technology Laboratories, NTT Corporation, Musashino-shi, 180-8585 Japan.
a) E-mail: lebreton.pierre.mz@hco.ntt.co.jp
b) E-mail: kazuhisa.yamagishi.vf@hco.ntt.co.jp

DOI: 10.1587/transcom.2019EBP3045

retraining of models for 10-sec sequences, and model transfer testing for long-term cases is discussed in Sect. 5. Section 6 addresses the applicability of long-term pooling modules to new models beyond the NTT and ITU-T P.1203 audiovisual-quality-estimation models, and Sect. 7 concludes this paper.

2. Related Work

2.1 Subjective Quality Characteristics

Research addressing the QoE of video-streaming services and how they relate to individual quality degradations, their relative importance, and temporal variation is presented below.

2.1.1 Individual Degradations

First, video-coding quality is positively correlated with QoE following a logistic function [16]–[18]. Decreasing resolution also results in lower QoE [17], [19]. However, the opposite is not necessarily true as higher resolution does not necessarily result in higher QoE since coding distortions can be larger [20]. As for frame-rate, frame-rate reduction decreases QoE, but the perception of it depends on the content's motion [21], [22]. Finally, stalling was found to be one of the strongest types of degradation [22]. It was found that long stalling events are better perceived than several short ones [23], and their position should also be considered [24], [25].

2.1.2 Relative Importance between Degradations

To provide a quantitative weighting between different types of distortions, previous studies have shown that initial loading delay is better perceived than stalling occurring in random positions in the video [26]. Users prefer images with coding distortions or blur due to spatial down sampling (also referred to as low image quality) to images without visible artifacts (also referred as high image quality) with stalling [27]. Similar results can be found with frame rate, as a lower frame rate is preferred to stalling [22]. It was found that for low-quality conditions, image quality is more important than temporal resolution [28]. Once image quality reached an appropriate level, it is temporal resolution that should be improved [29]. In terms of resolution, it was found that high image quality is preferred to high resolution at a low frame rate [16]. If only low throughput is available, a higher resolution with a lower frame rate should be used [30]. This comes with the exception of videos with a very small amount motion in which jerkiness can be easily identified [22]. Regarding temporal resolution, it was found that frame-rate reduction was better perceived than random frame drop [31]. The conclusion of these studies is that optimal QoE is not achieved by tuning individual axes (bit rate, frame rate, and resolution) separately, but in a global manner [32].

2.1.3 Temporal Aspects

So far in this paper, each degradation was constant over time. However, in adaptive bit-rate video streaming, videos are encoded at different resolutions, frame rates, and bit rates. Videos are split into small segments (or chunks), and a player can select which segment best matches the available throughput. This results in quality variation during playback. Several studies have addressed changes in video quality. It was found that adaptive bit-rate streaming improves user experience as it significantly decreases the number of stalling events [33]. However, this can only be achieved if the player optimally requests segments. Therefore, studies have addressed how players should behave to reach a high QoE. In these studies, it was found that increasing the video bit rate increases quality. However, users penalize decreasing quality more than reward increasing quality [34]. Quality changes should not occur too frequently, and having quality changes occurring more than once per second results in a lower experience than with constant low quality [35], [36]. Then, if the amplitude of a quality change is large, a switch to low quality is strongly penalized, but going to high quality will be well perceived [37]. In the case of low-amplitude changes, quality should be increased by maintaining resolution and frame rate before increasing the spatial and/or temporal resolution [35], [38]. Finally, the temporal-recency effect was also found to be important [35], [39].

2.2 Implications on Modeling

The general relationship between bit rate and QoE is frequently modeled using logistic regression [18], [40], [41]. However, bit-rate requirements are tightly interconnected with resolution. Therefore, while preliminary studies have defined different regression per resolution (or simply addressed only one resolution) to relate bit rate with video quality [40]–[42], more recent studies aimed at providing a unified model for estimating audiovisual quality using a continuous function of resolution [3], [4].

The integration of frame rate into an overall video-quality model was mostly found to be a multiplicative term with a baseline quality from coding [3], [4], [42], [43]. This allows the modeling of interactions between image quality and frame rate. Frame rate is mostly integrated in two different ways: 1) as an inverted falling exponential function for modeling jerkiness perception [4], [42], [43] and 2) as a combination with the number of pixels per frame enabling the modeling of the relationship between bit per pixel and mean opinion score (MOS) [3], [4], [40].

Previous studies have found that stalling should be considered as an additive term [44], [45], which has resulted in the general design of models evaluating the quality of the overall audio/visual sequence resulting from coding to which a factor that takes into account stalling is added [3], [4], [46]. The effect of stalling on quality is defined as a function of the stalling duration and position [3], [4], [27], [46].

The key challenges of adaptive bit-rate video-streaming-quality evaluation lies in the consideration of quality changes and how to aggregate them into an overall quality score. Pooling strategies based on Markov chain [47] or

auto-regressive [46] model have also been shown to be efficient. Alternatively, pooling with a per-second quality score weighted using an exponential decay giving more weight to recent quality scores was also found to be successful [4], [48]. Finally, it was found that even with a sophisticated pooling strategy, per-second quality estimation is of primary importance [49].

2.3 Extending the Scope to New Conditions

Based on these studies, accurate audiovisual-quality-estimation models have been proposed [4], [5], [9]. With the goal of real-time monitoring, meta-data-based models should be used as they are the least computing-intensive models. However, these models were designed to work under specific conditions: a given codec, given type of degradation, bit-rate range, frame-rate range, and resolution [4], [5]. These conditions are necessary as these models have only limited access to the video content and require several assumptions about how the videos are encoded. Extending the scope of these models to new conditions (4K-UHD and H.265/HEVC [50]) requires further study. The goal of this study is therefore to identify which parts of these models need to be retrained, identify new sets of coefficients, and investigate how previous models (including their coefficients) can be transferred to new conditions. In the following, the ITU-T P.1203 model in two different modes with different computational complexities (mode 0: meta-data-based, mode 3: bitstream-based) [5] and the NTT audiovisual-estimation model [4], [11] are considered. The second goal of this study is studying how the long-term pooling modules from these two models can be used together with other video-quality-estimation models (e.g., VMAF [13], BRISQUE [14], and NIQUE [15]). This analysis has two main outcomes. First, it allows the studying of the applicability of long-term modules to other video quality models than to those for which they were designed. Second, considering that adaptive bit-rate streaming-video quality estimation is still recent and only a few models are available that can support large quality variation and stalling events, this work shows how previous models designed for 10-sec sequences can be applied to these new scenarios would enable easy reuse of past efforts on video-quality modeling and increase the number of available models that support such conditions.

3. Model Description

This section gives an overview of the models addressed in this study. These models follow the same structure illustrated in Fig. 1. The meta-data-based models (NTT, ITU-T P.1203 mode 0) take as input meta-data about the audio and video streams (bit rate, frame rate, resolution) on a per-chunk basis, and the audio- and video-quality-estimation modules (O.21 and O.22, respectively) compute the audio- and video-quality scores on a per-second basis. The bit-stream model (ITU-T P.1203 mode 3) will, in addition to the data retrieved in mode 0, access the video bit stream to extract the quantization parameter providing video-quality scores with higher accuracy. The audiovisual-quality-estimation module (O.34) computes the audiovisual-quality score on a per-second basis using per-second audio- and video-quality scores. Then, temporal pooling of the per-second audiovisual-quality score is carried out in the quality-integration module (O.35) while taking into account the temporal effect. Finally, the last module (O.46) takes into account stalling events (start time and duration, both measured in seconds) that were reported by the video players and computes the overall quality score.

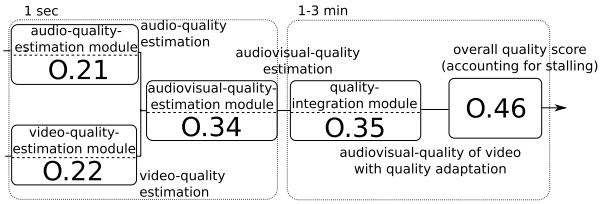

Fig. 1 General flow of audiovisual-quality estimation.

3.1 NTT Audiovisual-Quality-Estimation Model

The first model addressed was developed by Yamagishi et al. [4], [11] and is a meta-data-based no-reference model. Following the general flow described in Fig. 1, per-second video-only quality scores (from $O.22$) are computed using Eq. (1) with the video bit rate b_v, number of pixels per frame s, frame rate r, and constants v_{1-7}, which are the model's parameters. Please note that b_v, s, and r are time dependent. However, to simplify notation reference to time was removed on this variables. Finally, the equation is based on a logistic function, and quality increases along bit-rate with a saturation for high bit-rate values.

$$O.22(t) = X(t) + \frac{1 - X(t)}{1 + (\frac{b_v}{Y(t)})^{v_1}} \tag{1}$$

In the numerator, X denotes the maximum quality that can be reached considering the resolution and frame rate used.

$$X(t) = \frac{4 \times (1 - exp(-v_3 \times r)) \times s}{v_2 + s} + 1 \tag{2}$$

In the denominator, Y is a function of resolution and frame rate, which used together with bit rate, provides information on number of bits per pixel, and then provides information on quantization.

$$Y(t) = \frac{v_4 \times s + v_6 \times log_{10}(v_7 \times r + 1)}{1 - exp(-v_5 \times s)} \tag{3}$$

The audio-quality estimation module $O.21$ is commonly modeled using a logistic function:

$$O.21(t) = a_{1A} + \frac{1 - a_{1A}}{1 + (\frac{b_a}{a_{2A}})^{a_{3A}}} \tag{4}$$

The function depends on the audio bit rate b_a, and

a_{1A-3A} are constants that depend on the audio codec. b_a is time dependent, however, for simplicity of notation the reference to time was removed on this variable.

Audiovisual-quality estimation is based on audio and video quality with interaction terms (See Eq. (5)), as suggested in a previous study [51] (m_{1-4} are parameters of the model).

$$O.34(t) = m_1 + m_2 \times O.21(t) + m_3 \times O.22(t) + m_4 \times O.21(t) \times O.22(t) \quad (5)$$

Long-term aggregation is then carried out using Eq. (6). This equation allows the recency effect to be taken into account by providing a larger weight to recent quality scores, as proposed in a previous study [48]. The t_{1-5} are the model parameters, and T is the content duration. t refers to the time in seconds and $O.34(t)$ is the estimated audiovisual quality at time t. Therefore, the equation gives a higher weight to recent audiovisual quality scores than to older ones, enabling the recency effect to be modeled. Note that this equation makes use of two functions w_1 and w_2 used to simplify Eq. (6).

$$O.35 = \frac{\sum_{t=0}^{T} w_1(u) \times w_2(O.34(t)) \times O.34(t)}{\sum_{t=0}^{T} w_1(u) \times w_2(O.34(t))} \quad (6)$$

$$w_1(u) = t_1 + t_2 \times exp(\frac{u}{t_3})$$
$$w_2(O.34(t)) = t_4 - t_5 \times O.34(t) \quad (7)$$
$$u = \frac{t}{T}$$

Finally, the model takes into account stalling events using Eq. (8). The s_{1-3} are parameters of the model, n_b is the number of stalling events, t_b is the total duration of a stalling event, and a_b is the average duration between two stalling events (or 0 if no or one stalling event occurs). This allows the effect of long or too frequent stalling events on quality to be determined.

$$O.46 = 1 + (O.35 - 1) \times exp(-\frac{n_b}{s_1}) \times exp(-\frac{t_b}{T \times s_2}) \times exp(-\frac{a_b}{T \times s_3}) \quad (8)$$

3.2 The ITU-T P.1203 Model

The second model considered is ITU-T P.1203 [5]. It was designed to run in different modes to match different computational-complexity requirements. It ranges from mode 0 (a meta-data-based model) to mode 3 (a bitstream model). Modes 1 and 2 are intermediate complexity models. Between the different modes, the differences lie only in the computation of the video-quality-estimation scores from O.22, i.e., the video-quality estimations. Considering that the overall computational process of the ITU-T P.1203 model is complex and in depth details can be found in the respective ITU-T Recommendations P.1203 [5]–[8], only a general description is provided in this paper.

3.2.1 Video-Quality-Estimation Module: O.22

The first step is the computation of per-second video quality scores (from $O.22$). Only modes 0 and 3 are addressed as they are respectively the least computing intensive and most accurate. Across modes, the only difference is how quantization (Eq. (11)) is evaluated. Other aspects, such as frame-rate reduction and upscaling, are identical across modes. In mode 0, *quant* is defined as in Eq. (9) with *bpp* denoting the amount of bits per pixel (the bit rate divided by resolution and frame rate). In mode 3, *quant* is defined as in Eq. (10) with QP_{PB} denoting the average quantization parameter values for P and B frames. Note, that *quant* is a function of time, as bitrate and QP values change over time.

$$quant(t) = a_1 + a_2 \times ln(a_3 + ln(b_v) + ln(b_v \times bpp + a_4)) \quad (9)$$

$$quant(t) = \frac{\overline{QP_{PB}}}{51} \quad (10)$$

Then, it is possible to estimate the quality of the videos using Eq. (11).

$$M\hat{O}S_q(t) = q_1 + q_2 \times exp(q_3 \times quant(t)) \quad (11)$$

Here, $M\hat{O}S_q(t)$ denotes the estimated quality of videos affected by coding only at a time t. It is then converted to a degradation measure using Eq. (12) (RfromMOS defined in the E- Model [52]). The main idea of *RfromMOS/MOSfromR* is to address the non-linearity of the MOS scale.

$$D_q(t) = 100 - RfromMOS(M\hat{O}S_q(t)) \quad (12)$$

Two other degradations are considered in addition to quantization: upscaling and temporal. Upscaling degradation $D_u(t)$ relates display resolution with coding resolution at time t. While temporal degradation $D_t(t)$ measures the effect of frame-rate reduction on video quality at time t. This depends on three terms: one for frame-rate reduction itself, two masking terms related to interactions between frame rate and coding quality, and one for spatial resolution. Temporal degradation is non-null only when the frame rate is lower than 24 fps. After being linearly combined, a MOS estimate is obtained using the *MOSfromR* [52] (see Eq. (13)).

$$O.22(t) = MOSfromR(100-(D_q(t)+D_u(t)+D_t(t))) \quad (13)$$

When a mobile device is used, adjusting the visibility of the distortions due to screen size is necessary. Therefore, a three-order polynomial mapping is applied (Eq. (14)).

$$O.22_{mobile}(t) = \sum_{k=0}^{3} htv_k \times O.22(t)^k \quad (14)$$

3.2.2 Audio-Quality-Estimation Module: O.21

The O.21 module uses Eq. (15). Coefficients a_{1A}, a_{2A} and

Table 1 List of features used in the random forest model.

Feature
Initial stalling duration
Overall stalling duration
Stalling frequency
Ratio of stalling duration with regards to content length
Duration between the last stalling event and end of session
Average first third of $O.22$ scores (temporally ordered)
Average second third of $O.22$ scores (temporally ordered)
Average last third of $O.22$ scores (temporally ordered)
1st percentile of $O.22$ scores
5th percentile of $O.22$ scores
10th percentile of $O.22$ scores
Average first half of $O.21$ scores (temporally ordered)
Average second half of $O.21$ scores (temporally ordered)
Video duration

a_{3A} are codec dependent. A per-second audio MOS is found by applying the $MOSfromR$ function [52] to $QcodA$.

$$QcodA(t) = a_{1A} \times exp(a_{2A} \times b_a) + a_{3A} \quad (15)$$

3.2.3 Long-Term Aspects: Modules O.35/O.46

The long-term pooling and handling of stalling events of the ITU-T P.1203 model is based on a weighted average of an analytical and machine-learning pooling module. The analytical module computes audiovisual quality without stalling events (from $O.35$) using Eq. (16). In this equation, $O.35_{baseline}$ is the result of Eq. (6). The $negBias$ is an additional factor to account for large differences in quality during the session, $oscC$ is an additional term accounting for oscillations in audiovisual quality along the sequence, and $adaptC$ accounts for frequent large quality changes. Refer to ITU-T Rec. P.1203.3 [8] for details on the computation of these terms.

$$O.35 = O.35_{baseline} - negBias - oscC - adaptC \quad (16)$$

Once audiovisual quality is estimated (from $O.35$), the ITU-T P.1203 model takes into account the effect of stalling using Eq. (8). The final score is a weighted sum between the analytical pooling and outcome of a random forest model. The random forest model (described in ITU-T Rec. P.1203.3 [8]) is composed of 20 decision trees with a maximum depth of 6. It uses 14 features which are listed in Table 1.

4. Retraining of Short-Term Modules

The previous section introduced the computational models. This section describes the retraining and validation process to extend the models to new conditions.

4.1 Models Training

4.1.1 Design of Test Conditions

In this first evaluation eight videos (source reference circuit, SRCs) were used. These videos were native 3840x2160/60p or 7680x4320/60p shot using professional cameras. The videos with a resolution of 7680x4320 were converted to 3840x2160. Considering the limited number of SRCs used in the evaluation, special care was put into their selection. The selection process was done by covering four content-complexity scales: spatial information [53], temporal information [53], and motion complexity evaluated using average and variance of motion-vector norm (motion vectors obtained using an H.264 encoding at a high bit rate). Visual inspection was also done to ensure large variety of scenes and differences in colorfulness. The resulting selection was composed of a large variety of content, e.g., sports, festivals, landscapes, and theaters. Relationship between SRC number and type of content is listed in Table 2. Considering the goal to retrain the video-quality-estimation modules, 24 test conditions (hypothetical reference circuits (HRCs)) were considered. HRCs corresponding to processing of the video including coding, frame rate reduction, down-sampling, etc. These HRCs included resolutions from 240p to 2160p (with an aspect ratio of 16:9) with different frame rates and coding conditions (listed in Table 1). Only labels "REF", "HQ", "MQ", and "LQ" are given in the table as the coding conditions were defined per-source and per-resolution. This allows the content dependency to be addressed and ensures that each source was observed at low and high quality in a balanced manner. "REF" refers to the use of non-encoded reference video files, while the other three are coding conditions (high, medium, and low quality, respectively). To define these conditions, each SRC was encoded using the x265 codec preconfigured using the preset "slower" from FFmpeg[†]. A group of pictures (GOP) of two seconds was used. Encoding with several constant rate factor (CRF) values ranging from 21 to 44 were carried out allowing the bit-rate requirements per content and resolution to be identified. Videos were then encoded using a two-pass encoding with the CRF and a video-buffering verifier selected at per-source, per-resolution, and per-quality levels. This resulted in HQ having, on average, CRFs of around 24; MQ having, on average, a CRF around 32, and LQ having, on average,

Table 2 Details on SRCs used in the training phase

SRC	Type of content
01	Sport, bicycle race. Lots of motion and detailed textures.
02	Theater, a play. Little motion, colorful images, and detailed textures.
03	Landscape, landscape with trees. Little motion and detailed textures.
04	Festival, scene with people participating in a festival. Lots of people, complex motion, and detailed textures.
05	Landscape, fire burning trees. Complex motion and detailed textures.
06	Festival, scene with colorful kites moving. Colorful images, slow motion, and simple textures.
07	Landscape, aerial view of a field. Kids running, and wind moving leaves. Complex motion and detailed textures.
08	Sport, water polo match. Complex motion and detailed textures.

[†]https://ffmpeg.org

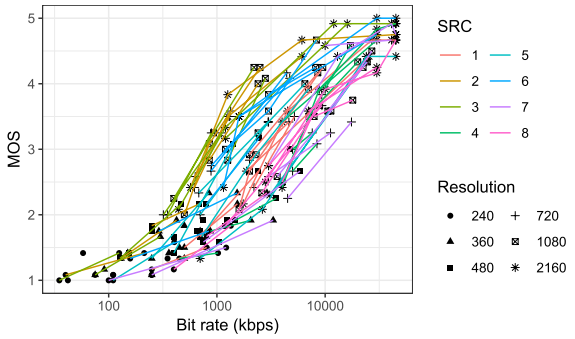

Fig. 2 Relationship between MOS and bit rate.

Table 3 Test condition matrix. The color in the cells allow matching frame rate and coding conditions.

Resolution	Frame rate		Quality levels					
2160	60	30	REF	HQ	MQ	LQ	HQ	MQ
1080	60	30	HQ	MQ	LQ		HQ	MQ
720	60	30	HQ		MQ		HQ	MQ
480	30	15	HQ		MQ		MQ	
360	30	15	HQ		MQ		MQ	
240	30	15	HQ		MQ		MQ	

a CRF of around 38. In terms of bit rate, these CRFs resulted in bit rates ranging from 40 kbps to 45 Mbps with an average of 5 Mbps. Detailed information is presented in Fig. 2. Once all content was encoded, the 192 processed video sequences (PVSs) were decoded using FFmpeg and stored in an uncompressed YUV format. If not encoded in 2160p, the PVSs were upscaled to this resolution using a lanczos3 algorithm [54]. If not at 60 fps, videos were converted to 60 fps by repeating frames. This allowed a well controlled upscaling technique to be used for playing back videos having different spatial and temporal resolutions.

4.1.2 Subjective Evaluation Procedure

The video-quality-evaluation procedure followed the absolute category rating (ACR) methodology using a five-point scale. The participants were asked in Japanese, "How would you rate the video quality?", the labels being: "5: Excellent", "4: Good", "3: Fair", "2: Poor", and "1: Bad". For this evaluation, 32 participants (16 males and 16 females, aged from 19 to 31: average 21) passed screening tests in visual acuity and color vision to participate. The choice of young persons was not by design, but resulting from the persons available during the hiring process. However, this is not expected to largely affect the results. Participants were non-experts with no previous experience in assessing audiovisual quality as part of their work. The playback was carried out using a professional-grade video player capable of playing back the uncompressed videos. A professional-grade 56-inch 4K-UHD TV was used. Participants were instructed to sit at 1.5H, (H the height of the screen). The room was a standardized laboratory environment designed for such evaluations: gray-wall room with controlled lighting [53]. The illumination was set to 20 lux, which corresponds to a dark room. After passing the vision tests, participants were provided written instructions about their task and went through a training phase in which six 10-sec PVSs were shown allowing them to practice their task. After this, the main task of the evaluation started. The main task was divided in three sessions of 64 PVSs. Each session was 20 min long. After each session, participants were given a rest. The presentation order of PVSs was randomized across participants.

Table 4 Coefficients for AVC and HEVC (video-quality-estimation module in NTT model).

Coefficient	H.264/AVC	H.265/HEVC	Scaling
v_1	1.8123	0.86802	**2.0878**
v_2	76116	341027	**0.22320**
v_3	0.11337	0.11461	0.989173
v_4	1.5371e-4	7.7064e-5	**1.9946**
v_5	0.99697	0.99697	1.0000
v_6	536.46	771.76	**0.69511**
v_7	0.14688	0.14688	1.0000

4.1.3 Subjective Evaluation Results

First, participants were found to have an inter-correlation higher than 0.9 (0.915 on average); hence, no participants were rejected. Figure 2 shows the relationship between bit rate and MOS. This graph shows the content dependency. For example, a bit rate of 1 Mbps resulted in a MOS of up to 3.5 for SRC 2, while the same bit rate resulted in a MOS of 1.5 for SRC 7.

4.1.4 Re-Training of Module *O*.22 in NTT Model

The retraining of the video-quality-estimation module O.22 (see Fig. 1) of the NTT model was achieved using non-linear regression. The optimizer nonlinear least squares (*nls*) from *R* was used. Initial values are of importance when performing non-linear regression, therefore initial values were set to those for the H.264-based encoded videos. This is motivated by the fact that optimal new coefficients for H.265-based encoded videos are expected to lies in a neighboring area in the search space to the coefficients for the H.264 case. As for the coefficients for H.264 encoded videos, these were obtained from the original author of this work, and are expected to be optimal [4], [11]. The new coefficients are listed in Table 4. The evolution of the coefficients between the AVC and HEVC conditions are of interest: the coefficient v_1, which takes into account the bit rate, was scaled by 2, which is in the same range as the expected improvement in coding efficiency of HEVC over AVC. The coefficient v_2, is an additive term with the number of pixels per frame, then if the number of pixels per frame is put into perspective, it can be observed that the coefficient v_2 did not largely vary and observed differences may only be due to the specific dataset under study. The next large change in coefficients is v_4, which was also scaled by 2. The interpretation of this change is complex as it can be justified by the increase in resolution interlinked with the change in coding efficiency of the HEVC codec. Finally, the last change is related to v_6, which may be related to the use

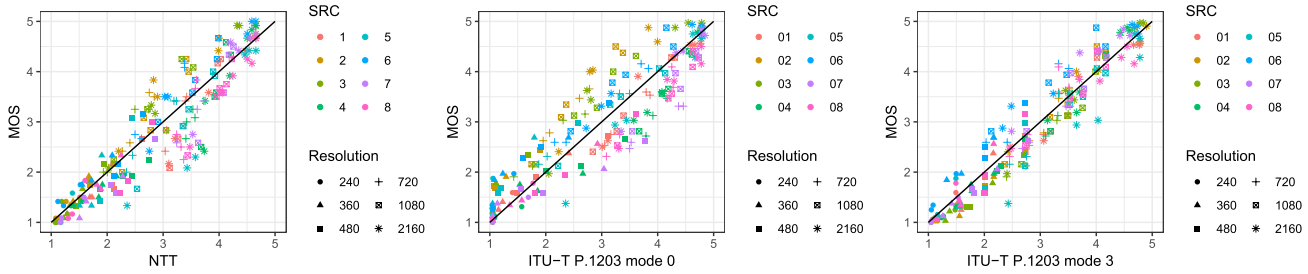

Fig. 3 Training performance - NTT (left), P.1203 mode 0 (center), mode 3 (right).

Table 5 Pearson correlation coefficient (PCC) and root mean square error (RMSE) in training. RMSE in bracket is obtained after a linear mapping.

Model	Former coef.		Retrained coef.	
	PCC	RMSE	PCC	RMSE
NTT	0.8256	1.0025 (0.6401)	0.9273	0.4482
ITU-T P.1203 m0	0.8361	1.0550 (0.6608)	0.9202	0.5077
ITU-T P.1203 m3	0.8646	1.293 (0.6050)	0.9623	0.3305

of 60-fps videos in this test.

Figure 3 (left) and Table 5 show the performance of this model at estimating the quality of the HEVC-encoded UHD videos using the new coefficients for H.265/HEVC. These results indicate that the model fit well the data. Validation of this model's performance is addressed in Sect. 4.2. Several conditions were overestimated for the high-resolution content: 2160p, 1080p, and 720p, as shown in Fig. 3 (left). This mostly involved SRCs 4, 5, and 8, which require a higher bit rate than the other SRCs to ensure high quality (see Fig. 2). The same phenomenon could be observed to a lesser extent for SRC 2 on the other side of the curve. This was expected as the NTT model is a meta-data-based model that does not take into account content dependency. Therefore, it estimates quality of a video with an average spatial and temporal complexity.

4.1.5 Retraining of Module $O.22$ in ITU-T P.1203 Model

The first step in retraining the video-quality-estimation module O.22 in the ITU-T P.1203 model involves identifying the new coefficients q_{1-3} in Eq. (11). Considering that the model handles the effect of upscaling from lower resolution and a lower frame rate than 24 fps at a later stage, it is necessary to consider only cases with native resolution and frame rate for retraining Eq. (11). Non-linear regression with initial values set to the AVC coefficients is carried out providing parameters q_{1-3}. Initial H.264 coefficients corresponding to the ones defined in the ITU-T Rec. P.1203.1 [6], and were obtained after a large effort of subjective testing involving many experiments performed by multiple laboratories [55]. Once found, these parameters are frozen as further steps of the optimization are carry out in the R domain, as defined in the E-Model [52]. After this process, coefficients u_{1-2} addressing the effect of spatial upscaling can be estimated using the subjective data of PVSs having a frame rate higher than 24 fps with any resolution. Once u_{1-2} are established, all conditions can be used to train coefficients that take into account the effect of frame-rate reduction and masking effect related to coding and upscaling. After going through all these different steps, the retrained ITU-T P.1203 model is obtained.

Figure3 (center) and (right) and Table 5 show the performance of the ITU-T P.1203 models retrained for HEVC. The retrained models fit the data well. The ITU-T P.1203 mode 3 model outperformed all the other models (ITU-T P.1203 mode 0, and NTT). This was expected as it is a bitstream model that allows addressing content dependency, while the other models do not. Table 6 and 7 provide the coefficients and their evolution between the training for AVC-encoded content and HEVC-encoded content. For the ITU-T P.1203 mode 0 model, coefficients a_{1-4} involved in Eq. (9) were approximately scaled by a factor of 2, as shown in Table 6. These coefficients relate bit rate and quality; therefore, this scaling seems consistent with expected coding-efficiency improvement of HEVC over AVC. Table 7 provides the new coefficients for the ITU-T P.1203 mode 3 model. It can be seen that the retraining mostly affected the coefficients q_{1-3} and u_{1-2}. This is to be expected as q_{1-3} reports on the differences of coding efficiency between AVC and HEVC. To compare AVC and HEVC coefficients, Eq. (11) is re-written in Eq. (17) to better reflect coefficients scaling (with $q_{1-3,avc}$, the original AVC coefficients). It can be observed that quantization parameters are shifted by a constant offset of 1.9131 (as the scaling of q_3 is added into the exponential function). Note that a change of QP by four units results in halving the bit rate. Therefore, there is a quarter offset between bit rate and quality compared with the AVC case. Moreover, the relationship between the decrease of bit rate and loss of quality is also slower than in the AVC case. This is an interesting result as it shows that with a given decrease of bit rate, there is a lower loss of quality for the HEVC case than for the AVC case. Finally, the changes in $u_{1,2}$ refer to the effect of downscaling on video quality. The change in these coefficients reflects the fact that in the previous training with HD and H.264, a MOS of 5 was given to the high quality HD videos. However, when 4K-UHD was introduced, this resulted in high quality HD videos receiving a new MOS value lower than 5 as the 5 rating is now given to the 4K-UHD videos. This stretching of the scale explains the change in coefficients $u_{1,2}$.

Table 6 Coefficients for AVC/HEVC (audio and video-quality-estimation modules in ITU-T P.1203 mode 0).

Coefficient	H.264/AVC	H.265/HEVC	Scaling
a_1	11.99835	6.554771	**1.830476**
a_2	-2.99992	-1.818259	**1.649886**
a_3	41.24751	19.017235	**2.168954**
a_4	0.13183	0.044088	**2.990156**
q_1	4.66	5.338358	0.8729276
q_2	-0.07	-0.230080	**0.304242**
q_3	4.06	2.865696	1.416759
u_1	72.61	62.381247	1.163972
u_2	0.32	0.043927	7.284813 ($\approx 10^{0.8624}$)
t_1	30.98	41.356120	0.7491032
t_2	1.29	1.29	1.0
t_3	64.65	64.65	1.0

Table 7 Coefficients for AVC/HEVC (video-quality-estimation module in ITU-T P.1203 mode 3).

Coefficient	H.264/AVC	H.265/HEVC	Scaling
q_1	4.66	5.007732	0.930561
q_2	-0.07	-0.010333	**6.7744**
q_3	4.06	5.620898	**0.72230**
u_1	72.61	42.270908	**1.7177**
u_2	0.32	0.496578	**0.64441** ($\approx 10^{-0.1908}$)
t_1	30.98	26.595857	1.16484
t_2	1.29	1.29	1.0
t_3	64.65	64.65	1.0

$$\begin{aligned} \hat{MOS}_q &= \frac{q_{1,avc}}{0.930561} + \frac{q_{2,avc}}{6.7744} \times exp(\frac{q_{3,avc}}{0.72230} \times quant) \\ &= \frac{q_{1,avc}}{0.930561} + exp(\frac{q_{3,avc}}{0.72230} \times quant - 1.9131) \end{aligned} \tag{17}$$

4.2 Validation of Model Performance

4.2.1 Design of Test Conditions

For validation, 32 SRCs not used in the training evaluation were used to evaluate content dependency. These SRCs show various scenes including sports, festivals, talk shows, theaters, landscape, and cooking (see Table 8). In terms of processing, the same combination of resolution and frame rate described in Table 3 was used. The encoding of each quality level was also done using x265 with the preset "slower". Instead of using an encoding using the CRF, it was decided to carry out a two-pass constant bit-rate encoding. The GOP size was increased from 2 to 4 sec. Finally, instead of choosing bit rates per PVS, it was decided to use the same bit rate across SRCs for a given combination of resolution and frame rate. Bit rates were chosen using the estimations from the retrained NTT model for balancing video quality in this evaluation. Figure 4 illustrates this approach. It shows a graph representing the relationship between between resolution, bit rate, frame rate, and quality. Then, bit rates values were chosen to distribute the conditions over the MOS scale. These are marked with solid black circles in Fig. 4.

Considering there were 32 SRCs and 24 HRCs, all combinations could not be tested in a reasonable amount of time. Therefore, HRCs were distributed over SRCs, and each SRC

Table 8 Details on SRCs used in the validation phase.

SRC	Type of content
01	Landscape, traffic on road, cars passing in from of the camera. Lateral slow motion and simple textures.
02	Waterfall, close-up on the water of a waterfall. Lots of complex motion and detailed textures.
03	Sport, video of a snowboarder going down a mountain. Lots of motion and simple textures.
04	Sport, athleticism: javelin throw. One person throwing a javelin. Little motion and complex textures in grass close-up.
05	Sport, athleticism: Hurdling. Five people competing at hurdling. Simple motion and simple textures.
06	Sport, bicycle race. Lots of motion and detailed textures.
07	Flower arrangement, people doing flower arrangement. Little to no motion, simple textures, and vivid color.
08	Flower arrangement, close-up on flower arrangement. Little to no motion and complex textures.
09	Flower arrangement, close-up on flower arrangement. Little to no motion, complex textures, and vivid color.
10	Sport, gymnastics: person running and posing. Static camera, motion limited to the gymnast. Simple textures.
11	Sport, gymnastics: person running and posing. Static camera, motion limited to the gymnast. Simple textures.
12	Landscape, fireworks: night scene with fireworks. Static camera, motion limited to the fireworks. Detailed textures on fireworks.
13	Sport, badminton: four players playing badminton. Far view, motion limited to players. Simple textures.
14	Sport, badminton: Close up of players playing badminton. Fast motion and simple textures.
15	Theater, a play. Little motion, colorful images, and detailed textures.
16	Festival, many people marching in line with flags, ornamental cars. Complex motion and textures.
17	Festival, view on a large number of people watching performances. Complex motion and textures.
18	Festival, far view on a large number of people running downhill. Complex motion and texture.
19	Landscape, corn fields with wind blowing leaves. Complex motion and very detailed textures.
20	Sport, outdoor scene with children jumping off a bridge into water. Complex motion in water and detailed textures.
21	Aerial view of a sailing ship on water. Simple motion and detailed textures on water and boat.
22	Close-up views of a steam train. Simple motion and detailed textures on steam and track.
23	Aerial view of Vienna. Motion limited to distant cars, complex texture with numerous buildings.
24	Landscape, view of a river with boat passing in front of the camera. City in the background. Little motion and simple textures.
25	Sport, people playing in parks. People running, birds flying, etc. Little motion and simple textures.
26	Time-lapse video of Tokyo. Little motion, complex texture.
27	Documentary, center of Vienna. Static camera filming horse-drawn carriage, pedestrian walking, and old buildings. Little motion and simple textures.
28	Dance, cameras filming many people waltzing. Complex motion and average texture complexity.
29	Cooking, close-up of cooking. Little to no motion, colorful images, and very detailed textures.
30	Cooking, close-up of cook preparing and cutting fish. Little to no motion, colorful images, and very detailed textures.
31	Sport, far view on a water polo match. Complex motion and detailed textures.
32	Sport, close-up view on a water polo match. Complex motion and detailed textures.

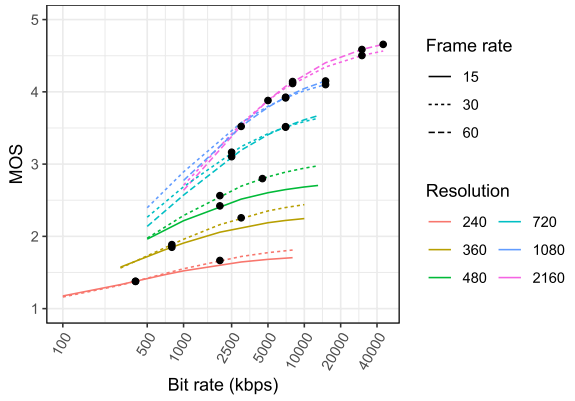

Fig. 4 Relationship between MOS and bit rate. Estimations from NTT model. Solid black circles mark selected coding conditions.

was processed only six times, resulting in a total of 192 PVSs. Special care was taken during the distribution of HRCs across SRCs; therefore, the SRCs were equally shown in both high and low quality.

Finally, six PVSs of this validation were replaced with 6 PVSs from the former training evaluation to study the correlation between the subjective rating across evaluations.

4.2.2 Subjective Evaluation Procedure

This subjective evaluation was identical to that of the training described in Sect. 4.1.2. Thirty-two different participants (16 male, 16 female, aged from 18 to 23; average 20.8) participated in this evaluation. One participant largely deviated from the others with an average inter-participant correlation of 74%; the others maintained an average inter-participant correlation of about 90%. Using the common set of 6 PVSs, agreement between the data collected during the training and validation evaluations (hereafter, training and validation datasets, respectively) could be studied. Correlation was high: Pearson correlation coefficient (PCC) = 0.978 and root means square error (RMSE) = 0.260, and the two datasets had the following linear relationship: $MOS_{training} = 0.911 \times MOS_{validation} + 0.3705$.

4.2.3 Performance Evaluation

Figure 5 and Table 9 show the performance accuracies of the three models on the validation dataset. The retrained coefficients were evaluated on this dataset without updating any coefficients. Each model performed well on this dataset. The ITU-T P.1203 mode 3 model was found to be highly correlated with the subjective data, in terms of correlation it outperforms the others models thanks to its bitstream approach allowing it to address content dependency. However, when the RMSE is considered, it can be seen that the model generally over-estimates the quality of the videos. The differences between RMSE with and without linear mapping show that the estimations are highly correlated, but with a constant offset of about 0.9 MOS. An in-depth analysis was performed and revealed that the encoding with constant rate factor (CRF) (used for the training dataset) resulted in a distribution of QP values in P and B-frames different from that for the constant bit rate encoding (CBR) (used in the validation dataset). Results show that the relationship between average bit rate and quality was kept constant, as proved by the similar performance of both NTT and ITU-T P.1203 mode 0 models, which only use average bit rate values to estimate the effect of coding on picture quality. However, the change in the type of encoding resulted in higher QP values for P and B frames in the training dataset than in the validation dataset. Since the ITU-T P.1203 mode 3 model only uses the average QP for P and B frames to estimate the effect of coding on quality (discarding information on QP for I-frames, average bit rate, etc.), the model over-estimated the quality of videos encoded using CBR. This shows a limit in the usage of the ITU-T P.1203 mode 3 model. Finally, it can be seen that the ITU-T P.1203 mode 0 and NTT models performed well.

To further evaluate the models, a cross validation was conducted: each model was retrained on the validation dataset and tested on the training dataset. The RMSEs of 0.4304, 0.4264, and 0.3911 for the NTT and ITU-T P.1203 mode 0 and 3 models, were respectively obtained during training. While on validation, the RMSEs were 0.4495, 0.5732, and 0.4231 for the NTT and ITU-T P.1203 modes 0 and 3 models, were respectively obtained. This shows that for the NTT and ITU-T P.1203 mode 3 models, carrying out the inversion of training and verification datasets does not significantly alter the model's performance. For the ITU-T P.1203 mode 0 model, however, there was a large drop in performance. The problem may be related to the fact that the validation dataset contains many SRCs and few repetitions of the same source with different coding conditions making training more difficult for that model considering the required data partitioning in the training process.

5. Long-Term Aspects and Model Transfer

The previous section addressed the retraining of the video-quality-estimation module, $O.22$. In this section, testing the assumption that long-term modules ($O.35$/$O.46$) can be transferred to new conditions is discussed.

5.1 Design of Test Conditions

For testing long-term modules, four different subjective evaluations were conducted. Two using 1-min SRCs and two using 3-min SRCs (each duration having one test using a TV, and one using a smartphone (mobile)). The SRCs selected in these evaluations showed a large variety of scenes shot in Japan. The content was similar to common TV shows in Japan and depicted scenes such as scenery, festivals, different sporting events, documentaries, interviews, etc. Information on SRC can be found in Table 11. The evaluations with 1-min SRCs involved 30 SRCs and 30 HRCs. Each SRC was processed with only two HRCs, resulting in 60 PVSs. In the evaluations with 3-min SRCs, 11 SRCs and 22 HRCs

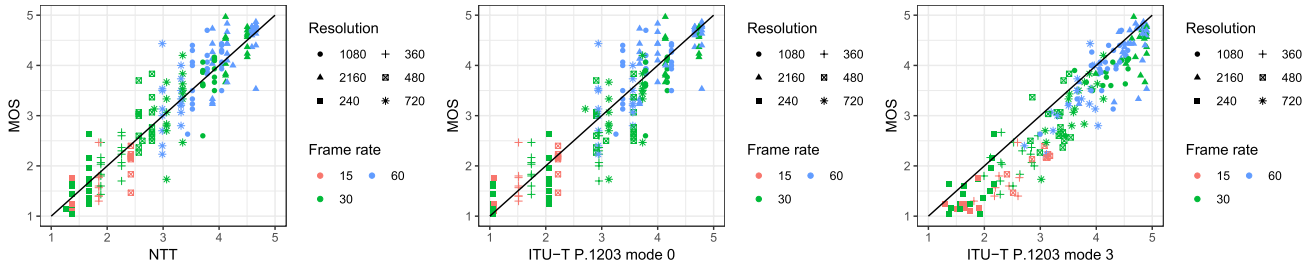

Fig. 5 Validation performance: NTT (left), P.1203 mode 0 (center), mode 3 (right).

Table 9 PCCs, Spearman correlation coefficients (SRCCs), and RMSEs in validation set. RMSEs in validation dataset. RMSEs in parentheses were obtained after linear mapping. RMSE/Algn used common dataset to align training and verification datasets.

Model	PCC	SRCC	RMSE	RMSE/Algn
NTT	0.9306	0.9194	0.4169 (0.3773)	0.3977
P.1203 mode 0	0.9081	0.8891	0.4817 (0.4317)	0.4604
P.1203 mode 3	0.9428	0.9344	0.5309 (0.3428)	0.4490

Table 10 Quality levels for both TV and mobile (M). Res. indicates spatial resolution, B.R. is bit rate in kbps, FR is frame rate. GOP indicates GOP size in seconds.

QL	Res. (TV)	B.R. (TV)	Res. (M)	B.R. (M)	FR	GOP	Audio B.R.
Q0	240	1000	144	100	30	2	64
Q1	360	1500	360	450	30	2	96
Q2	480	3000	480	640	30	2	96
Q3	720	5000	720	1000	60	2	128
Q4	2160	14000	2160	4000	60	2	128
Q5	240	1000	144	100	30	1	64
Q6	360	1500	360	450	30	1	96
Q7	480	3000	480	640	30	1	96
Q8	720	5000	720	1000	60	1	128
Q9	1080	10000	1080	4000	60	1	128
Q10	2160	20000	2160	8000	60	1	196

were used resulting in 22 PVSs. In each case, the processing involved quality adaptation and simulating stalling events. Each segment was encoded using x265, using two pass encoding, constant bit rate, and preset "slower". In terms of GOP size, either 1 or 2 sec was chosen depending on the HRC. Considering that bit-rate requirements differ for mobile devices and TVs, bit rates were distinct between visualization devices. Figure 6 provides an overview of the test conditions. The first number indicates the quality level and the value in parentheses refers to the duration of the condition. The white boxes indicate stalling events, which were simulated by freezing the video with a dynamic loading wheel. The details about each quality level for both mobile and TV are given in Table 10. Finally, audio was encoded using an AAC-LC audio codec at the bit rates listed in Table 10. These conditions were selected on the basis of the goal of retraining the long-term audiovisual quality estimations models to support new conditions. Achieving this goal requires the use of a large span of test conditions with various combinations of coding conditions and stalling events. Therefore, test conditions need to be designed with various quality adaptation and stalling events.

5.2 Subjective Evaluation Procedure

The subjective evaluation was similar to the previous evaluations: the same TV, laboratory environment with controlled conditions (gray-wall room, illumination at 20 lux, viewing distance at 1.5H, etc.), subjective scales, methodology, etc. were used. The main differences lie in the use of audiovisual content instead of content with video only. Special care was taken so each PVS was listened to at 73-dB SPL using headphones. In the mobile evaluations, a 5.5-inch 3840×2160 smartphone was used, and the viewing distance was 5–7 H. Similarly, participants were asked to evaluate the overall quality of the videos on a 5 grade scale using the ACR methodology. The evaluations were divided in 15-min sessions with breaks between sessions.

After screening, 4×32 participants (4×16 male, and 4×16 female, ages from 18 to 37; average 21) took part in the evaluations. Table 12 shows the inter-participant agreements across all different evaluations. Outlier rejection was conducted by rejecting participants who had a lower Pearson correlation coefficient than 0.7 compared to the mean rating. Going from 10-sec video sequences to long video sequences (1 and 3 min) resulted in a significant decrease in inter-participant agreement from 90 to 80%. This can be explained by long video sequences needing higher complexity than short ones in order to be evaluated. Indeed, in the case of short videos, only few quality changes occur. Therefore, it is easy for the participants to provide an overall rating. On the other hand, in the case of long videos, lots of quality changes occur. Therefore, the participants need to remember the quality of the video throughout the sequence and decide what their overall experience was, which may be challenging. In addition, note that long videos introduce a new type of degradation: stalling events. The impact of stalling events on the overall quality can also add extra difficulty to the task of the participants, as they need to identify how much it annoyed them compared with the other degradation. All of this resulted in larger variation across participants. This is supported by the data as a larger agreement across participants was observed in the 3-min evaluations than in the 1-min evaluations for which more frequent quality variations were visible (see Fig. 6), hence making the overall quality integration more difficult for the participants. Finally, a similar inter-participant correlation was observed when comparing TV and mobile.

Table 11 Details on SRCs used in the long-term video quality tests. SRC_1 and SRC_3 are respectively the SRCs used in the 1 and 3 min tests.

SRC_1	SRC_3	Type of content
01	01	Documentary, interviews and presentation of antiques in an old house. Little motion and detailed textures.
02	02	Documentary, Snow Festival in Hokkaido. Little motion and detailed textures: falling snow and large crowds.
03		Documentary, illuminations at night. Little motion and simple textures.
04	03	Interviews, people walking around old Japanese streets. Little motion and detailed textures.
05		Documentary, museums about Sherlock Holmes. Old buildings. Little motion and detailed textures.
06		Documentary, visit to a China Town in Japan. Person walking in the streets. Little motion, detailed textures, and colorful images.
07	04	Sport, presentation and interview of players in an indoor sports event. Average motion and little texture.
08		Sport, presentation and interview of players in an indoor sports event. Average motion and little texture.
09		Gymnastics, gymnast practicing on parallel bars and horizontal bars. Motion limited to the gymnast. Little texture.
10		Festival, many people marching in line with flags, ornamental cars. Complex motion and textures.
11		Documentary on winter sports resort. Little motion and detailed textures on trees, snow, and crowds.
12	06	Sport, camera following snowboarders down a mountain. Lots of motion and detailed texture on trees and snow.
13	07	Documentary about flyboard (a water sport). Complex motion and detailed texture due to water motion.
14		Documentary about flyboard (a water sport). Complex motion and detailed texture due to water motion.
15	08	Documentary about drum festival in Japan. Motion limited to performers and detailed textures on trees, grass and costumes.
16		Music scene: people singing in a church. Little to no motion, simple texture, and colorful video.
17	09	Cooking show about pies. Little to no motion, simple texture, and colorful video.
18		Cooking show about Mexican food. Little to no motion, simple texture, and colorful video.
19	10	Cooking show about Japanese food. Little to no motion, simple texture, and colorful video.
20		Video about athleticism. Sports include javelin throw, hurdling, pole vault. Motion limited to athletes, detailed textures.
21		Festival, view of a large number of people watching performances. Complex motion and textures.
22	11	Sport, gymnastics: person running and posing. Static camera, motion limited to the athlete. Simple textures.
23		Landscape: mountains with lots of trees and snow falling. Little to no motion and very detailed texture.
24		People playing badminton. Far and close-up views. Motion limited to players. Simple textures.
25		Theater, a play. Little motion, colorful images, and detailed textures.
26	05	Travel documentary. Contains both aerial and close-up views. Little motion and detailed textures.
27		Scenery video. Aerial views of fields and rivers. Very complex motion due to wind on leaves and highly detailed textures.
28		Documentary, center of Vienna. Static camera filming streets followed by scene with people waltzing. Various motion and texture complexity.
29		Cooking, close-up of cooking: cutting fish and vegetables. Little to no motion, colorful images, very detailed textures.
30		Sport, far and close-up views of a water polo match. Complex motion and detailed texture.

5.3 Performance Evaluation and Retraining

Thanks to the modular design of the models, the extension of quality estimations per 10 sec up to 3 min would in principle not require further retraining as the long-term pooling modules use scores in the MOS domain. Therefore, previous training of the long-term temporal aggregation performed in the AVC and HD cases is expected to be transferable to HEVC and 4K-UHD. This section tests this assumption and compares it with a full retraining of the models.

Figure 7 shows the performance of the different models on the 1- and 3-min TV and mobile datasets. The quantitative results are listed in Table 13. The transferred coefficients refer to the use of the coefficients of long-term integration

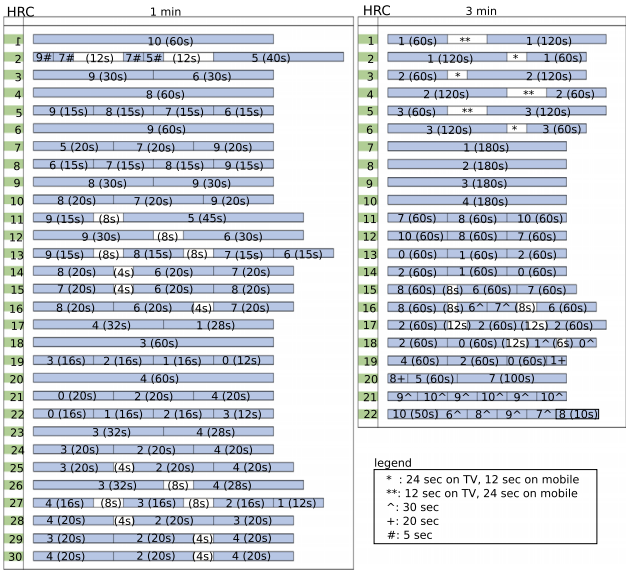

Fig. 6 Quality adaptation in long-term evaluations.

Table 12 Inter-participant agreement across evaluations.

Display	Dur.	# Inliers	# Outliers	Avg. R	Std. R
TV	10 sec	32	0	0.9150	0.02497
TV	10 sec	32	0	0.8979	0.04256
TV	1 min	29	3	0.7906	0.05033
TV	3 min	27	5	0.8289	0.07516
Mobile	1 min	24	8	0.7916	0.07524
Mobile	3 min	27	5	0.8343	0.07086

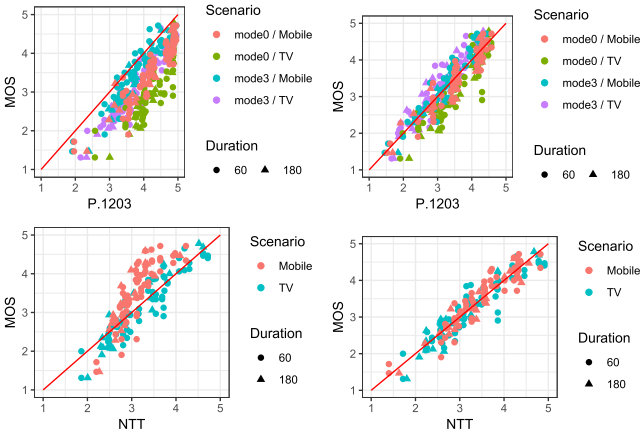

Fig. 7 Performance results. ITU-T P.1203 (top), NTT (bottom). Transferred coefficients (left), retrained coefficients (right).

Table 13 Performance evaluation. M0 and M3 correspond to use of ITU-T P.1203 modes 0 and 3 models, respectively. RMSEs in parentheses were obtained after linear fitting. Dur. refers to content duration.

Dur.	Model	TV		Mobile	
		PCC	RMSE	PCC	RMSE
Transferred coefficients, retrained O.22 module					
1 min	NTT	0.9288	0.3000 (0.2821)	0.8275	0.6256 (0.4277)
3 min	NTT	0.9470	0.3244 (0.2865)	0.8524	0.6023 (0.4220)
1 min	M0	0.8440	1.1754 (0.4085)	0.9200	0.7226 (0.2985)
3 min	M0	0.9365	1.2679 (0.3128)	0.9349	0.7581 (0.2863)
1 min	M3	0.9266	0.7480 (0.2864)	0.9336	0.3821 (0.2728)
3 min	M3	0.9816	0.6115 (0.1700)	0.9518	0.4945 (0.2475)
Transferred coefficients, linear mapping on retrained O.22 module					
1 min	NTT	0.9283	0.3487 (0.2833)	0.8271	0.5820 (0.4281)
3 min	NTT	0.9467	0.3220 (0.2873)	0.8514	0.5648 (0.4232)
1 min	M0	0.8964	0.5496 (0.3376)	0.9290	0.3118 (0.2818)
3 min	M0	0.9729	0.6082 (0.2061)	0.9431	0.3082 (0.2683)
1 min	M3	0.9287	0.3331 (0.2825)	0.9232	0.4590 (0.2927)
3 min	M3	0.9750	0.3614 (0.1981)	0.9317	0.3711 (0.2930)
Retrained coefficients, retrained O.22 module					
1 min	NTT	0.9232	0.3071 (0.2926)	0.9342	0.2754 (0.2717)
3 min	NTT	0.9574	0.2736 (0.2575)	0.9635	0.2192 (0.2158)
1 min	M0	0.8842	0.4642 (0.3558)	0.9422	0.2663 (0.2550)
3 min	M0	0.9687	0.5506 (0.2213)	0.9296	0.3227 (0.2973)
1 min	M3	0.9293	0.3118 (0.2812)	0.9578	0.2612 (0.2189)
3 min	M3	0.9448	0.3796 (0.2923)	0.9616	0.2349 (0.2214)

modules ($O.35/O.46$ in Fig. 1) obtained for AVC/HD and applied to HEVC/4K-UHD. Without retraining or applying any type of mapping, these modules from the NTT and ITU-T P.1203 models achieved reasonable performance. The NTT model achieved high performance for TV. However, a non-linear relationship between estimations and ground truth data was observed for the mobile. For the ITU-T P.1203 model, the estimations appeared linearly correlated with the ground truth data in each mode and each dataset (data collected from the 1-min TV, 3-min TV, 1-min mobile, and 3-min mobile evaluations). However, the model generally over-estimated video quality. This over-estimation was larger in the mode 0 model compared to the mode 3 model, resulting in the saturation of quality estimations.

To address this saturation problem, a linear mapping of the video quality scores at a per-second basis (from $O.22$ in Fig. 1) before the pooling module is needed. To identify this linear mapping, a unique pair of coefficients (a,b) such as $O.22_{lm} = a \times O.22 + b$ was estimated such that the values of a and b are constant across the four datasets and for the two considered modes (0, and 3). The optimization was carried out using the GRG non-linear engine solver from Microsoft Excel, enabling a single pair of coefficients (a, b) to be found while minimizing the sum of RMSEs across all modes and datasets. For fairness of comparison, a second pair of coefficients (a, b) was also estimated for the NTT model. The results are listed in Table 13. Applying this mapping enabled the over-estimation and saturation issues of the ITU-T P.1203 model to be addressed. For the NTT model, applying the linear mapping to the per-second video-quality score had limited effect as estimations were already aligned for TV, and the mobile required a non-linear mapping. It should be noted that linear mapping has not always been beneficial to the NTT model, as performances on the TV slightly decreased, but the loss in performance for TV was translated into performance improvement for mobile.

Table 14 Coefficients for AVC and HEVC (quality-integration module in NTT model).

Coefficient	H.264/AVC TV	H.264/AVC Mobile		H.265/HEVC TV & Mobile		
a_{1A}	4.36209	4.36209	a_{1A}	4.36209	htv_1	−7.81834
a_{2A}	16.4606	16.4606	a_{2A}	16.4606	htv_2	11.9270
a_{3A}	2.08184	2.08184	a_{3A}	4.36209	htv_3	−4.02027
m_1	0.620119	1.757568216	m_1	0.000	htv_4	0.44680
m_2	0	0.00910769	m_2	0.151201		
m_3	0.613691	0.002708346	m_3	0.000018		
m_4	0.068487	0.133572238	m_4	0.217927		
t_1	0.006666	0.013031751	t_1	0.0106366		
t_2	4.04E-05	2.18252E-06	t_2	0.000026287		
t_3	0.156498	0.10372705	t_3	0.145071		
t_4	0.14318	0.147889458	t_4	0.0140164		
t_5	0.023864	0.024168639	t_5	0.002354253		
s_1	11.35587	9.963211795	s_1	4.2040		
s_2	6.140927	19.12417144	s_2	4593.696154		
s_3	3.932605	7.850157023	s_3	4.84229		

Table 15 Coefficients for AVC and HEVC (quality-integration module in ITU-T P.1203 model).

Coefficient	H.264/AVC	H.265/HEVC	Coefficient	H.264/AVC	H.265/HEVC
a_{1A}	100.0	100.0	c_2	7.85416481	10.7476813
a_{2A}	−0.05	−0.05	c_{23}	0.0185382	0.052042006
a_{3A}	14.60	14.60	ct_1	0.6775608	0.70273296
m_1	−0.001	0.000	ct_2	−8.05533303	0
m_2	0.153743	0.202116	ct_3	0.17332553	4.557959273
m_3	0.971539	0.000000	ct_4	−0.01035647	0.052336362
m_4	0.024618	0.164773	s_1	9.3516	4.1596
t_1	0.0066662	0.0000410	s_2	0.918908	14.154886
t_2	0.000040402	0.000000006	s_3	11.05676	2.05361
t_3	0.156498	11.680155	htv_1	−0.60293	−5.523305279
t_4	0.1431797	0.1274659	htv_2	2.12382	8.377083675
t_5	0.023864156	0.026448231	htv_3	−0.36936	−2.374009707
c_1	1.87403625	5.102600956	htv_4	0.03409	0.224675601

To study the improvement due to a full retraining of the models, new coefficients for the long-term-aggregation modules $O.35/O.46$ of the NTT and ITU-T P.1203 models was estimated. Tables 14 and 15 list the old and new coefficients after retraining for the NTT and ITU-T P.1203 models, respectively.

The long-term-aggregation modules $O.35/O.46$ of the ITU-T P.1203 model were designed to be independent of modes and visualization devices. Therefore, a single set of coefficients is used across modes (0–3) and devices (TV, mobile). The differences between modes and devices are addressed with the video-quality-estimation module O.22. Therefore, the retraining process of modules $O.35/O.46$ is carried out jointly across all datasets and modes. The datasets containing videos of different durations (1 and 3 min) resulted in a different number of PVSs per dataset (60 in the 1-min evaluations and 22 in the 3-min evaluations). Optimizing the RMSE by joining datasets (across 82 PVSs/point) can be problematic as the retraining of coefficients would be mostly driven by the 1-min evaluations since they contain more PVSs. To address this issue, the RMSE is computed per dataset (across 60 and 22 PVSs separately), and the optimization process is carried out to decrease the sum of the two RMSEs. By doing this, equal weights are

given to the 1- and 3-min evaluations. In this study, we aimed at optimizing the long-term pooling module for modes 0 and 3. Therefore, the overall objective was to decrease the sum of RMSE computed for each dataset and each mode. This result in the sum of 8 RMSE values: 4 for each dataset using modes 3 and 4 for each dataset using mode 0. This optimization was carried out using the GRG non-linear engine solver from Microsoft Excel.

Among the four datasets, two datasets contained mobile video tests. In the ITU-T P.1203 model, the handling of the differences between mobile and TV scenarios is handled by a third-order polynomial function (see Eq. (14)), which maps video quality scores, in $O.22$, estimated for TV to mobile. This mapping belongs to the module $O.22$. However, the $O.22$ module re-training discussed in Sect. 4.1.4 did not include mobile video data; thus, this polynomial mapping could not be re-trained. The identification of the coefficients for the mapping between TV and mobile (htv_{1-4} in Eq. (14)) was done jointly with the coefficient for the model. This was achieved using an iterative process: first the coefficients for $O.35/O.46$ were estimated using the TV datasets only. Optimization was done jointly across modes 0 and 3. The initial coefficients were set to the original values given in ITU-T Recommendation P.1203 for the AVC/HD case. Once a new set of coefficients was found, coefficients for $O.35/O.46$ were frozen, and optimization was done on the mobile datasets across modes 0 and 3. This enabled identifying coefficients htv_{1-4}, which allowed mapping between the TV and mobile subjective ratings. Once coefficients htv_{1-4} were identified, they were frozen and the coefficients of modules $O.35/O.46$ were fine-tuned using all datasets (mobile/TV, 1 min/3 min) and modes (0/3). This process was then repeated iteratively until convergence was reached. In the original design of the NTT model, the model had two sets of coefficients for O.35/O.46: one for TV and one for mobile. For fair comparison across models, it was decided to use a similar approach for the retraining as for the ITU-T P.1203 model. Therefore, only a single set of coefficients was used for both mobile and TV. A three-order polynomial function was also applied to the per-second video-quality estimate from O.22 to map estimations to mobile ratings. The overall training process was similar to that for the ITU-T P.1203 model. The final coefficients are listed in Table 14.

The performances of the retrained models are summarized in Table 13. As expected, retraining improved the performance of all models. The ranking in terms of performance slightly varied among the datasets, nevertheless the bitstream-based ITU-T P.1203 mode 3 model generally outperformed the meta-data-based models. Statistical significance of the differences among the models was tested based on RMSE obtained after linear fitting using an F-test with a 95% confidence level, as described in ITU-T Rec. P.1401 [56]. The results indicate that, without retraining, i.e., transferred coefficients, the ITU-T P.1203 in mode 3 model significantly outperformed the NTT and ITU-T P.1203 in mode 0 models. In this case, the performances of the NTT and ITU-T P.1203 mode 0 models were found statistically equivalent. After retraining, all models were statistically equivalent. However, with a larger number of datasets, it is expected that the ITU-T P.1203 in mode 3 model would significantly outperform others as having more data makes statistical significance more easily reachable and could allow better training of the models, as described in the cross validation section (Sect. 4.2), as distribution of conditions in training dataset can be critical when training a model.

6. Use of Long-Term Modules with Other Video-Quality-Estimation Models

In this section, evaluation of long-term-integration modules $O.35/O.46$ with a large variety of video-quality-estimation modules is discussed. There are two reasons for this evaluation. First, it allows the robustness of the long-term pooling module to be tested by showing its independence to a specific video-quality-estimation module. Second, it allows the extension of current video-quality-estimation models to new conditions for which they are not designed, e.g., large quality variation and stalling events. Conducting such an evaluation is important considering that adaptive bit-rate video streaming is widely used and only a few models have been designed to address this scenario. Therefore, showing how previous studies on quality estimation of 10-sec videos can be extended to new scenarios is of great importance.

Therefore, the performance of the long-term pooling modules from the NTT and ITU-T P.1203 models with models designed for the quality estimation of traditional 10-sec videos was studied. In this analysis, several models were compared: NIQUE [15], GMSD [57], PSNR, PSNR-HVS [58], VIFP [59], SSIM [60], VMAF [13], BRISQUE [14], and MS-SSIM [61]. Considering that each model has its own scale (for example, VMAF $\in [0, 100]$, SSIM $\in [0, 1]$, etc.) and the relationship between scores from the models and MOS is non-linear, an appropriate mapping to the MOS scale needs to be identified. Therefore, the dataset composed of 10-sec video sequences described in Sect. 4.1 was used. This allowed a fair comparison with the retrained models presented in this paper, as they were also trained on this dataset. Two mapping function were considered: a sigmoid function ($f(x) = \frac{A}{C+exp(B\times x)}$) and exponential function ($f(x) = A+B\times exp(C\times x)$) as it was found that the most appropriate mapping function differs across models (suitable mapping functions can be found in Table 16). These functions were chosen because they are common approaches for mapping estimations from the algorithm to subjective data [62]. Then, a key challenge with using the 10 sec validation and training datasets previously introduced is the handling of frame-rate reduction. In these datasets, videos are encoded with various frame rates (as listed in Table 3). This is problematic as none of the considered models were designed to address frame-rate reduction. Therefore, two strategies are listed in Table 16: "MF" and "AF". "MF" (matching frames) corresponds to restricting the analysis to PVS having the same frame rate as the source (60 fps). "AF" (all frames) corresponds to using all PVSs and repeats frames to have the

Table 16 Performance evaluation. Pooling 1 & 2 uses ITU-T P.1203 or the NTT long-term pooling modules. MF/AF correspond to different handling of frame-rate reduction.

Model	Mapping	MF						AF					
		10 sec		1+3 min Pooling 1		1+3 min Pooling 2		10 sec		1+3 min Pooling 1		1+3 min Pooling 2	
		PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE
TV													
NIQUE* [15]	Sigmoid	0.2020	0.9372	0.8252	0.4581	0.8151	0.4706	0.6334	0.9320	0.8365	0.4475	0.8281	0.4589
GMSD [57]	Sigmoid	0.5225	0.8159	0.6630	0.6282	0.7180	0.5811	0.4166	1.0950	0.6693	0.6237	0.7255	0.5750
PSNR	Exponential	0.4478	0.8556	0.7814	0.5218	0.7811	0.5182	0.4354	1.0843	0.7929	0.5101	0.7910	0.5084
PSNR-HVS [58]	Exponential	0.5724	0.7847	0.8037	0.4973	0.8145	0.4820	0.4873	1.0518	0.8181	0.4824	0.8270	0.4691
VIFP [59]	Sigmoid	0.7366	0.6472	0.7389	0.5605	0.7370	0.5595	0.6447	0.9207	0.7541	0.5468	0.7495	0.5490
SSIM [60]	Exponential	0.4701	0.8446	0.6689	0.6211	0.6715	0.6186	0.3967	1.1057	0.6735	0.6175	0.6757	0.6153
VMAF [13]	Sigmoid	0.9012	0.4147	0.8628	0.4242	0.8831	0.3915	0.7220	0.8333	0.8813	0.3983	0.8989	0.3685
BRISQUE* [14]	Sigmoid	0.4929	0.8326	0.7630	0.5261	0.7730	0.5111	0.7711	0.7668	0.7841	0.5083	0.7930	0.4951
MS-SSIM [61]	Exponential	0.7268	0.6572	0.7471	0.5605	0.7693	0.5365	0.5116	1.0349	0.7648	0.5434	0.7826	0.5227
P.1203 mode 3*	-	-	-	-	-	-	-	0.9384	0.3911	**0.9370**	**0.2867**	0.9474	0.2647
P.1203 mode 0*	-	-	-	-	-	-	-	0.9264	0.4264	**0.9264**	**0.2885**	0.9323	0.2929
NTT*	-	-	-	-	-	-	-	0.9342	0.4304	0.9353	0.2945	**0.9403**	**0.2750**
Mobile													
NIQUE* [15]	Sigmoid	0.2020	0.9372	0.7844	0.4933	0.7878	0.4913	0.6334	0.9320	0.7911	0.4877	0.7957	0.4846
GMSD [57]	Sigmoid	0.5225	0.8159	0.6793	0.5851	0.7035	0.5660	0.4166	1.0950	0.6781	0.5873	0.7031	0.5675
PSNR	Exponential	0.4478	0.8556	0.7569	0.5234	0.7469	0.5306	0.4354	1.0843	0.7647	0.5158	0.7549	0.5232
PSNR-HVS [58]	Exponential	0.5724	0.7847	0.7622	0.5171	0.7519	0.5260	0.4873	1.0518	0.7685	0.5122	0.7583	0.5212
VIFP [59]	Sigmoid	0.7366	0.6472	0.7469	0.5301	0.7350	0.5380	0.6447	0.9207	0.7557	0.5220	0.7439	0.5305
SSIM [60]	Exponential	0.4701	0.8446	0.6751	0.5841	0.6738	0.5866	0.3967	1.1057	0.6793	0.5806	0.6777	0.5835
VMAF [13]	Sigmoid	0.9012	0.4147	0.8106	0.4727	0.8200	0.4618	0.7220	0.8333	0.8175	0.4644	0.8258	0.4553
BRISQUE* [14]	Sigmoid	0.4929	0.8326	0.7401	0.5397	0.7532	0.5259	0.7711	0.7668	0.7534	0.5288	0.7667	0.5150
MS-SSIM [61]	Exponential	0.7268	0.6572	0.7203	0.5596	0.7321	0.5464	0.5116	1.0349	0.7351	0.5470	0.7450	0.5350
P.1203 mode 3*	-	-	-	-	-	-	-	0.9384	0.3911	**0.9597**	**0.2201**	0.8807	0.3591
P.1203 mode 0*	-	-	-	-	-	-	-	0.9264	0.4264	**0.9359**	**0.2761**	0.9370	0.2810
NTT*	-	-	-	-	-	-	-	0.9342	0.4304	0.9171	0.4650	**0.9488**	**0.2437**

same number of frames as the source. No-reference models (marked with "*" in Table 16) performs better when all PVSs were considered, while full-reference models performs better when only PVSs with the same frame rate as the source were used. This is to be expected, as full-reference models need to compare matching frames or they would report large distortions and no-reference models do not have this constrain. Then, if all PVSs are considered, no-reference models have more data points allowing them to achieve a more robust fit as being less impacted by each estimation error around each point. Therefore, the mapping function relating model scores and MOS is trained using only PVSs at 60 fps for the full-reference metrics and all PVSs for the no-reference metrics.

To address long-term videos, two different long-term pooling strategies were considered: the one from ITU-T P.1203 and that from the NTT model (referred to as Pooling 1 and Pooling 2 in Table 16). The general approach is to execute each model on a per-frame basis, then scores of audio and video quality are averaged on a per-second basis giving 60 audio ($O.21$) and video ($O.22$) scores in the 1-min evaluations, and 180 audio and video scores in the 3-min evaluations. Then, per-second scores are provided to the long-term pooling modules that perform temporal aggregation and take stalling events into account.

Similarly to the 10-sec video case, special care is needed to address segments having 30 fps instead of 60 fps. Therefore, Table 16 shows the two handling strategies: "MF" only compares matching frames with the reference. While the second strategy "AF" repeats frames of the segments with 30 fps to reach 60 fps and compares frame-by-frame with the reference. This table also shows the performance of both strategies.

During the model comparison, VMAF [13] showed highly competitive performance and was the best (ITU-T P.1203 and NTT models put aside). Its performance was higher for TV compared to mobile. The results of NIQUE [15] (a non- reference, pixel-based model) indicate the importance of long-term pooling since its performance on the 10-sec training and validation datasets was low, but performed well in the long-term (1 and 3 min) datasets. This difference in performance can be justified by the fact that long-term pooling accounts for stalling, which is a significant degradation. It also aggregates quality estimations over long periods, decreasing the effect of per-second estimation errors. Finally, models other than the ITU-T P.1203 and NTT models, performed generally better for TV compared to mobile.

The three models retrained in this study are in bold in Table 16. The performance of the ITU-T P.1203 and NTT models were higher than the other models. However, it should be stressed that these models were only designed for evaluating videos encoded with H.265/HEVC (or H.264/AVC with the initial training), while the other models are more flexible. A model such as VMAF, which is pixel-based, allows a wider variety of codecs and degradation types to be handled. However, this comes with the cost of being more computationally intensive (not to mention that it is also

a full-reference model). Therefore, there is a trade-off between computational complexity and the need for supported conditions.

7. Conclusion

Six subjective evaluations involving 192 participants were conducted to evaluate the quality of videos encoded with H.265 in adaptive bit-rate video-streaming scenarios under various conditions (resolution, frame rate, duration, and devices). Based on the collected data, the main goal was to retrain two state-of-the-art audiovisual-quality-estimation models designed for adaptive bit-rate video streaming: the NTT model [4], [11] and ITU-T P.1203 in (modes 0 and 3) [5]–[8]. The robustness of the models was addressed by looking into the changes in coefficients between the pre-existing H.264 coefficients and new H.265 coefficients, and it was found that changes in coefficients reflected the expected improvement in coding performance of HEVC over AVC. Moreover, the paper highlighted the validity of the framework's structure as long-term integration modules could be transferred to new conditions without requiring retraining. The fact that coefficients trained for HD using AVC could be transferred to HEVC and 4K-UHD without further effort is an important result as it indicates that the support of new conditions such as 8K- UHD or other video codecs, such as AV1 or VP9, is possible with limited subjective testing: only the evaluations with 10-sec videos would be needed.

Another outcome of the datasets from all six evaluations was the ability to report on the participants' agreement and reliability across different conditions: short-term evaluations (validation and training) allowed higher inter-participant agreement compared to long-term evaluations (1 and 3 min), and fewer outliers were found for TV compared to mobile. Another contribution of this paper was its comparison of various models. This comparison validated the high performance of the NTT and ITU-T P.1203 models but also the applicability of long-term aggregation modules to new models. With this approach, VMAF [13] could be extended to new scenarios for which it was not designed. Finally, with this model comparison, we showed the importance of long-term pooling and handling of stalling features. Even models with average performance in the short-term datasets could perform well in the long-term datasets. All these results are believed to be of high interest as they have extended the support of current models and indicate to which extent these models can be transferred to new conditions. Future work will involve model transfer, addressing new evaluation concepts, and investigating how previous studies on QoE estimation can be applied to estimate user engagement and viewing time.

References

[1] K. Brunnström, S.A. Beker, K.D. Moor, A. Dooms, and S. Egger, et al., "Qualinet white paper on definitions of quality of experience," European Network on Quality of Experience in Multimedia Systems and Services, 2012.

[2] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," IEEE Commun. Surveys Tuts., vol.17, no.1, pp.469–492, 2015.

[3] W. Robitza, M.N. Garcia, and A. Raake, "A modular HTTP adaptive streaming QoE model - Candidate for ITU-T P.1203 ("P.NATS")," QoMEX, 2017.

[4] K. Yamagishi and T. Hayashi, "Parametric quality-estimation model for adaptive-bitrate-streaming services," IEEE Trans. Multimedia, vol.19, no.7, pp.1545–1557, 2017.

[5] ITU-T Rec. P.1203, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport," 2017.

[6] ITU-T Rec. P.1203.1, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – video quality estimation module," 2017.

[7] ITU-T Rec. P.1203.2, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – audio quality estimation module," 2017.

[8] ITU-T Rec. P.1203.3, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – quality integration module," 2017.

[9] C.G. Bampis and A.C. Bovik, "Learning to predict streaming video QoE: Distortions, rebuffering and memory," arXiv:1703.00633, 2017.

[10] ITU-R Rec. BT.2020, "Parameter values for ultra-high definition television systems for production and international programme exchange," ITU-R, 2015.

[11] K. Yamagishi, "Audio-visual quality extimation device, method for estimating audio-visual quality and program," US Patent 15/776425, 2018.

[12] ITU-T Rec. H.265, "High efficiency video coding," 2018.

[13] Z. Li, A. Aaron, L. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Netflix Technology Blog, 2016.

[14] A. Mittal, A.K. Moorthy, and A.C. Bovik, "No-reference image quality assessment in the spatial domain," IEEE Trans. Image Process., vol.21, no.12, pp.4695–4708, 2012.

[15] A. Mittal, R. Soundararajan, and A.C. Bovik, "Making a completely blind image quality analyzer," Signal Process. Lett., vol.20, no.3, pp.209–212, 2013.

[16] D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield, "Toward optimal rate control: A study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate," Proc. SPIE 5150, Visual Communications and Image Processing, 2003.

[17] M. Cheon and J. Lee, "Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience," IEEE Trans. Circuits Syst. Video Technol., vol.28, no.7, pp.1467–1480, 2018.

[18] K. Yamagishi and T. Hayashi, "Parametric packet-layer model for monitoring video quality of IPTV services," IEEE Int. Conf. on Communications (ICC), p.110–114, 2008.

[19] J. Li, Y. Koudota, M. Barkowsky, H. Primon, and P.L. Callet, "Comparing upscaling algorithms from HD to Ultra HD by evaluating preference of experience," QoMEX, 2014.

[20] P.L. Callet, S. Pechard, S. Tourancheau, A. Ninassi, and D. Barba, "Towards the next generation of video and image quality metrics: Impact of display, resolution, contents and visual attention in subjective assessment," IMQA, 2007.

[21] G. Ghinea and J.P. Thomas, "QoS impact on user perception and understanding of multimedia video clips," Proc. 6th ACM Int. Conf. on Multimedia, pp.49–54, 1998.

[22] Q. Huynh-Thu and M. Ghanbari, "Temporal aspect of perceived quality in mobile video broadcasting," IEEE Trans. Broadcast., vol.54, no.3, pp.641–651, 2008.

[23] Y. Qi and M. Dai, "The effect of frame freezing and frame skipping on video quality," Proc. Int. Conf. on Intelligent Information Hiding and Multimedia, pp.423–426, 2006.
[24] T. Minhas and M. Fiedler, "Impact of disturbance locations on video quality of experience," Workshop QoEMCS, 2011.
[25] T. Hoßfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial delay vs. interruptions: Between the devil and the deep blue sea," QoMEX, 2012.
[26] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," IEEE Trans. Multimedia, vol.10, no.7, pp.1316–1324, 2008.
[27] K. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of experience estimation for adaptive HTTP/TCP video streaming using H.264/AVC," IEEE CCNC, 2012.
[28] K. Yamagishi and T. Hayashi, "Video quality planning model for videophone services," IEEE Globecom, 2006.
[29] R. Rajendran, M.V.D. Schaar, and S.F. Chang, "FGS+: Optimizing the joint SNR–temporal video quality in MPEG-4 fine grained scalable coding," IEEE International Symposium on Circuits and Systems (ISCAS), 2002.
[30] J. Lee, F.D. Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," IEEE Trans. Multimedia, vol.13, no.5, pp.882–893, 2011.
[31] R.R. Pastrana-Vidal, J.C. Gicquel, C. Colomes, and H. Cherifi, "Frame dropping effects on user quality perception," Int. WIAMIS, 2004.
[32] N. Cranley, P. Perry, and L. Murphy, "User perception of adapting video quality," Int. J. Hum.-Comput. St., vol.64, no.8, pp.637–647, 2006.
[33] J. Yao, S. Kanhere, I. Hossain, and M. Hassan, "Empirical evaluation of HTTP adaptive streaming under vehicular mobility," Networking, pp.92–105, 2011.
[34] B. Lewcio, B. Belmudez, A. Mehmood, M. Wältermann, and S. Möller, "Video quality in next generation mobile networks-perception of time-varying transmission," Workshop on Communications Quality and Reliability, 2011.
[35] M. Zink, J. Schmitt, and R. Steinmetz, "Layer-encoded video in scalable adaptive streaming," IEEE Trans. Multimedia, vol.7, no.1, pp.75–84, 2005.
[36] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen, "Flicker effects in adaptive video streaming to handheld devices," Proc. 19th ACM Int. Conf. on Multimedia, 2011.
[37] M. Grafl and C. Timmerer, "Representation switch smoothing for adaptive HTTP streaming," PQS, pp.178–183, 2013.
[38] A. Moorthy, L. Choi, A. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," IEEE J. Sel. Topics Signal Process., vol.6, no.6, pp.652–671, 2012.
[39] T. Hoßfeld, M. Seufert, C. Sieber, and T. Zinner, "Assessing effect sizes of influence factors towards a QoE model for HTTP adaptive streaming," Workshop QoMEX, 2014.
[40] M. Garcia, P. List, S. Argyropoulos, D. Lindegren, M. Pettersson, B. Feiten, J. Gustafsson, and A. Raake, "Parametric model for audiovisual quality assessment in IPTV: ITU-T Rec. P.1201.2," MMSP, 2013.
[41] Y. Shen, Y. Liu, Q. Liu, and D. Yang, "A method of QoE evaluation for adaptive streaming based on bitrate distribution," Int. Conf. on Communications (ICC), 2014.
[42] J. Joskowicz and J. Ardao, "Combining the effects of frame rate, bit rate, display size and video content in a parametric video quality model," LANC, 2011.
[43] Y.F. Ou, T. Liu, Z. Zhao, Z. Ma, and Y. Wang, "Modeling the impact of frame rate on perceptual quality of video," IEEE Int. Conf. on Image Processing (ICIP), 2008.
[44] M. Garcia, D. Dytko, and A. Raake, "Quality impact due to initial loading, stalling and video bitrate in progressive download video services," QoMEX, 2014.
[45] M.N. Garcia, F.D. Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnström, and A. Raake, "Quality of experience and HTTP adaptive streaming: A review of subjective studies," QoMEX, 2015.
[46] Z. Duanmu, A. Rehman, K. Zeng, and Z. Wang, "Quality of experience prediction for streaming video," ICME, 2016.
[47] H. Yeganeh, R. Kordasiewicz, M. Gallant, D. Ghadiyaram, and A.C. Bovik, "Delivery quality score model for Internet video," IEEE Int. Conf. on Image Processing (ICIP), 2014.
[48] T. Hayashi, G. Kawaguti, J. Okamoto, and A. Takahashi, "Subjective quality estimation model for video streaming services with dynamic bit-rate control," IECE Trans. Commun., vol.E89-B, no.2, pp.297–303, Feb. 2006.
[49] M.N. Garcia, W. Robitza, and A. Raake, "On the accuracy of short-term quality models for long-term quality prediction," QoMEX, 2015.
[50] G.J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," IEEE Trans. Circuits Syst. Video Technol., vol.22, no.12, pp.1649–1668, 2012.
[51] M.H. Pinson, W. Ingram, and A. Webster, "Audiovisual quality components," IEEE Signal Process. Mag., vol.28, no.6, pp.60–67, 2011.
[52] ITU-T Rec. G.107, "The E-model: a computational model for use in transmission planning," ITU-T, 2015.
[53] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," ITU-T, 2008.
[54] K. Turkowski and S. Gabriel, "Filters for common resampling tasks," Graphics Gems I, A.S. Glassner, ed., pp.147–165, Academic Press, 1990.
[55] A. Raake, M.N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on, 2017.
[56] ITU-T Rec. P.1401, "Statistical analysis, evaluation and reporting guidelines of quality measurements methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," 2012.
[57] W. Xue, L. Zhang, X. Mou, and A.C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," IEEE Trans. Image Process., vol.23, no.2, pp.684–695, 2014.
[58] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," International Workshop on Video Processing and Quality Metrics, 2006.
[59] H. Sheikh and A. Bovik, "Image information and visual quality," IEEE Trans. Image Process., vol.15, no.2, pp.430–444, 2006.
[60] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol.13, no.4, pp.600–612, 2004.
[61] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," IEEE Asilomar Conference on Signals, pp.1398–1402, 2003.
[62] ITU-R Rec. BT-500-13, "Methodology for the subjective assessment of the quality of television pictures," ITU-R, 2012.

Pierre Lebreton received his Engineering degree in Computer Science from Polytech' Nantes, France in 2009. In 2010 he joined the group Assessment of IP-based Applications at the Technical University of Berlin, where he studied towards his PhD on QoE and depth perception of 3D stereoscopic videos. After graduating, he joined the group Audio Visual Technology at the Technical University of Ilmenau in 2015 and studied image aesthetic appeal and QoE of video-streaming services through crowdsourcing experiments and big-data analysis. In 2016, he joined the group Networked Sensing and Control at Zhejiang University, China, where he applied data analysis to the Bi-cycle Sharing System and industrial control systems. In December 2017, he joined NTT Laboratories and now focuses his research on quality and user-engagement prediction for video-streaming applications.

Kazuhisa Yamagishi received his B.E. degree in Electrical Engineering from the Tokyo University of Science in 2001 and his M.E. and Ph.D. degrees in Electronics, Information, and Communication Engineering from Waseda University in Japan in 2003 and 2013. Since joining NTT Laboratories in 2003, he has been engaged in the development of objective quality-estimation models for multi-media telecommunications. From 2010 to 2011, he was a visiting researcher at Arizona State University. He received the Young Investigators' Award (IEICE) in Japan in 2007, the Telecommunication Advancement Foundation Award in Japan in 2008, the ITU-AJ Encouragement Award in 2017, and the TTC Award for distinguished service in 2018.