PAPER

Bitstream-Quality-Estimation Model for Tile-Based VR Video Streaming Services

Masanori KOIKE[†a)], **Yuichiro URATA**[†b)], ***and*** **Kazuhisa YAMAGISHI**[†c)], ***Members***

SUMMARY Tile-based virtual reality (VR) video consists of high-resolution tiles that are displayed in accordance with the users' viewing directions and a low-resolution tile that is the entire VR video and displayed when users change their viewing directions. Whether users perceive quality degradation when watching tile-based VR video depends on high-resolution tile size, the quality of high- and low-resolution tiles, and network condition. The display time of low-resolution tile (hereafter delay) affects users' perceived quality because longer delay makes users watch the low-resolution tiles longer. Since these degradations of low-resolution tiles markedly affect users' perceived quality, these points have to be considered in the quality-estimation model. Therefore, we propose a bitstream-quality-estimation model for tile-based VR video streaming services and investigate the effect of bitstream parameters and delay on tile-based VR video quality. Subjective experiments on several videos of different qualities and a comparison between other video quality-estimation models were conducted. In this paper, we prove that the proposed model can improve the quality-estimation accuracy by using the high- and low-resolution tiles' quantization parameters, resolution, framerate, and delay. Subjective experimental results show that the proposed model can estimate the quality of tile-based VR video more accurately than other video quality-estimation models.

key words: *tile-based VR, subjective experiment, bitstream-quality-estimation model*

1. Introduction

Virtual reality (VR) video is expected to become more common due to the increased resolution of recent head-mounted displays (HMDs) [1], the increase in coding efficiencies [2], [3], and the widening of networks (e.g., 5G and fiber optics). However, since a high bitrate is required to provide high-quality VR video, network traffic is increasing. To prevent generating extremely high bitrates, VR videos need to be encoded at a suitable bitrate. To do that, a quality-estimation model needs to be developed to determine whether VR video achieves high quality.

To reduce network traffic in VR video streaming, standardization organizations discussed tile-based VR video streaming. Omnidirectional MediA Format (OMAF) has been developed as a tile-based VR video streaming method [4]. In this streaming, VR video is divided into grid-based tiles and each tile is encoded at multiple bitrates. User's viewing direction tiles are distributed at high bitrates and other tiles are distributed at low bitrates or are not distributed. Another method of tile-based VR video streaming was proposed [5], [6], as shown in Fig. 1. In this streaming, one high-resolution tile and one low-resolution tile are distributed from the server. The low-resolution tile represents the entire VR video and is distributed constantly. Divided tiles from the entire VR video are encoded at a high resolution, where high-resolution tiles are slightly overlapped. The user's viewing direction is sent from the HMD, and the server sends a high-resolution tile of the user's viewing direction to the HMD. To reduce quality degradation, the method for optimizing the bitrate of tile-based VR video distribution in a wireless network is proposed [7], [8]. Zhao et al. [7] reduce the re-buffering time by adjusting the coding rate for each GOP and the transmission rate for each transmission slot. Guo et al. [8] optimized the transmission time and power allocation of base station or access point as well as the encoding rate of each tile to maximize the video quality.

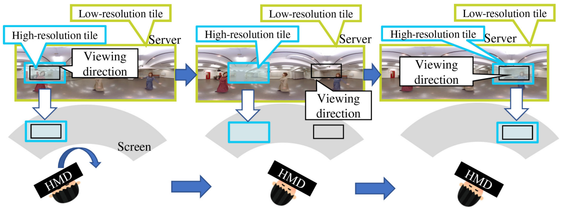


Fig. 1 Tile-based VR video streaming.

Since the high-resolution tiles are basically displayed on the HMD, users mainly watch a high-quality video. However, when users change their viewing directions, they perceive the upscaling degradations of the low-resolution tiles because the low-resolution tiles are displayed on their HMD [9]. In addition, the perceived quality depends on the display time of the low-resolution tiles (hereafter, delay) [10], [11]. Therefore, the coding artifacts on both high- and low-resolution tiles, the upscaling degradations of low-resolution tiles, and delay need to be taken into account in quality-estimation models.

The quality-estimation models are classified into full reference (FR), reduced reference (RR), and no reference (NR) models on the basis of their input. The FR pixel-based model takes the original and degraded images as input and can evaluate the impact of video codec and contents on the quality because it uses both original and decoded images [12]–[14]. Since the FR pixel-based model uses the video signals, the model does not need to be optimized/trained per

Manuscript received September 29, 2021.
Manuscript revised January 11, 2022.
Manuscript publicized February 18, 2022.

[†]The authors are with the Network Service Systems Laboratories, NTT Corporation, Musashino-shi, 180-8585 Japan.
a) E-mail: masanori.koike.wx@hco.ntt.co.jp
b) E-mail: yuuichirou.urata.kg@hco.ntt.co.jp
c) E-mail: kazuhisa.yamagishi.vf@hco.ntt.co.jp

DOI: 10.1587/transcom.2021EBP3163

codec [15]. However, the FR pixel-based model needs a large amount of computational power because original and degraded videos are used. The RR pixel-based model takes features of the original images and degraded images [16], [17]. The RR pixel-based model does not need to be optimized per codec. The RR pixel-based model uses a lower amount of computational power than the FR pixel-based model because it uses features of the original images that have less information than the original video. NR pixel-based models only take the degraded images and evaluate the video quality. However, NR pixel-based models have not been standardized due to the lack of quality-estimation accuracy [18], [19]. In addition, because of the absence of source information, NR pixel-based models are usually less accurate than the corresponding FR and RR counterparts [15]. On the one hand, NR bitstream-quality-estimation models have been standardized with a good quality-estimation accuracy [20], [21]. The accuracy of quality estimation is ensured by switching the set of coefficients of the model for different codecs because NR bitstream-quality-estimation models need to be optimized per codec. Since the NR bitstream-quality-estimation model only needs to parse bitstream, it can estimate the quality with a small amount of computational power.

To select a suitable bitrate for the tile-based VR video streaming, it is important to measure quality at the head end, in which the encoder and streaming server are located [22]. In this case, the amount of calculation has to be lower because so many videos are encoded at the same time.

In this paper, to monitor the tile-based VR video at the head end, a bitstream-quality-estimation model is proposed. To take into account the coding artifacts on both high- and low-resolution tiles and delay, the upscaling degradations of low-resolution tiles, quantization parameter (QP), resolution of high- and low-resolution tiles, and delay are used as input of the model. To develop and validate the model, subjective experiments are carried out. To verify the quality-estimation accuracy in detail, the proposed model is compared with several conventional models in terms of accuracy.

The remainder of this paper is organized as follows. Section 2 provides quality factors about VR videos and related work about the quality-estimation models of videos. Section 3 explains our proposed bitstream-quality-estimation model. Section 4 details the method, conditions, and procedure of conducted subjective experiments. Section 5 describes the results of experiments and the estimation accuracy of the proposed model. Finally, Sect. 6 concludes our paper.

2. Related Work

To develop a quality-estimation model for tile-based VR video, the features/parameters to be used as input for the model need to be determined. To determine the parameters, quality-influencing factors have to be clarified. Next, since many quality-estimation models have been proposed for 2D videos that may be extended to tile-based VR video, these models have to be investigated. In addition, the issues that need to be addressed in existing tile-based VR video quality-estimation models should be summarized.

2.1 Quality-Influencing Factors

To develop a quality-estimation model of tile-based VR videos, the quality-influencing factors have to be clarified. In 2D video streaming, the relationship between encoded video quality and many quality-influencing factors (e.g., bitrate, framerate, resolution, QP, codec [23], and search range of motion vector [24]) is investigated, and the quality of videos and bitrate in several conditions are shown.

Similar to 2D video quality, many VR quality-influencing factors are also investigated. Duan et al. [25] showed the effect of the combination of bitrate, framerate, and resolution on VR video quality, and Tran et al. [26] showed similar characteristics by varying the combination of QP and resolution. Yang et al. [27] showed the effect of framerate on VR video quality, and Han et al. [28] showed the effects of bitrate and resolution. From the perspective of a network, Fei et al. [29] showed the effects of bitrate, network latency, and packet loss. Singla et al. [30] showed the effect of resolution, bandwidth, network round-trip latency, and bitrate on VR video quality. The effect of quality degradation on tile-based VR video has been reported [31], [32], and the display time of low-resolution tiles has been shown to affect quality. Van der Hooft et al. [33] indicated the quality degradation due to switching between viewport tiles. Han et al. [34] indicated the quality degradation due to stalling and quality switching in adaptive bitrate streaming, and Fuente et al. [35] showed the effect of display time of the low-resolution tiles on tile-based VR video quality.

As discussed above, many quality-influencing factors of VR videos are clarified. To monitor the quality of encoded tile-based VR videos at the head end, the parameters from the bitstream of the tile-based VR videos need to be used, e.g., bitrate, framerate, resolution, and QP. Those quality-influencing factors must be considered to make a quality-estimation model that has high accuracy.

2.2 Quality-Estimation Models

Many quality-estimation models have been proposed in 2D [20], [21], [36]–[41] and VR videos [42]–[47].

To estimate the 2D video quality, bitstream-quality-estimation models have been proposed. Takagi et al. [36] used the inverse of the exponential function of QP, Izumi et al. [37] used the exponential function of QP, and Keimel et al. [38] used QP with a motion vector. Anegekuh et al. [39] used QP and content information extracted from motion vectors. Also, the Laplacian mixture probability density function [40], discrete cosine transform (DCT) coefficients [41], and other parameters are used to estimate 2D video quality. ITU-T standardized a bitstream-quality-estimation model for H.264/AVC up to HD video (i.e., the P.1203.1 mode 3 model [20]) and for H.265/HEVC and VP9 up to 4K video (i.e., the P.1204.3 model [21]). However, these models aim to estimate the quality of 2D videos and cannot be used for estimating the tile-based VR video quality because they do not consider the impact of low-resolution tiles on quality and delay.

To estimate VR video quality, several studies have been conducted. Fremerey et al. [42] described a quality-estimation model using bitrate, framerate, and resolution and found that it estimated quality with high accuracy. Machine learning methods are also used to estimate VR videos. Anwar et al. [43] used QP and Kim et al. [44] used resolution and other parameters related to quality of video for input to machine learning and proposed quality-estimation models with machine learning. These papers [42]–[44] show quality-estimation models for VR videos that are not tile-based VR and are encoded uniformly, so they did not aim to estimate the tile-based VR video or consider the degradation caused by low-resolution tiles. Li et al. [45] showed the quality-estimation model of tile-based VR over a wireless network with bitrate, resolution, and other network parameters (i.e., stalling or bitrate level switching). R. Schatz et al. [46] estimated tile-based VR video quality using weighted peak signal-to-noise ratio (wPSNR), structural similarity index measure (SSIM), and video multimethod assessment fusion (VMAF). These papers show the quality-estimation model of tile-based VR video, but the degradation caused by low-resolution tiles and delay is not considered. Shaowei et al. developed a model that takes into account the effects of delay and low-resolution tiles [47]. They showed the tile-based video quality-estimation model that uses the low-resolution tile quality degradation with quantization stepsize, spatial resolution and refinement duration.

In the existing tile-based VR video quality-estimation model, the degradation caused by low-resolution tiles is not always taken into account. While changing from low-resolution tiles to high-resolution tiles, the delay occurs, and users perceive quality degradation. To take low-resolution tiles' quality into account, delay needs to be taken into consideration because delay affects users' length of time to watch low-resolution tiles. Therefore, to estimate high- and low-resolution tiles and to take low-resolution tiles' quality into account, we estimate tile-based VR video quality with high- and low-resolution tiles' QP, resolution, and delay.

2.3 Comparative Models

Among the quality-estimation models described in Sect. 2.2, promising models that can be extended to tile-based VR video are described in detail. We choose three bitstream-quality-estimation models [21], [39], [47], which are used to estimate video quality using bitstream.

The first model is that of Anegekuh et al. [39]. This model shows a 2D video quality-estimation model using bits per frame and non-zero motion vectors. In this model, temporal complexity is extracted from the ratio of non-zero motion vectors, and spatial complexity is extracted from QP and the number of bits of B and I frames. In each content, content type (CT) is calculated by this temporal and spatial complexity. CT is calculated as follows:

$$CT = \left(\left(\frac{1}{L} \sum_{i=1}^{L} \frac{bits_{\mathrm{Ii}}}{Max_{\mathrm{bits}}} \right) \times \left(\frac{1}{S} \sum_{x=1}^{S} \frac{bits_{\mathrm{Px}}}{Max_{\mathrm{bits}}} \times \frac{1}{QP_{\mathrm{Ii}}} \right) \right)^{\alpha} \times \left(\frac{1}{M} \sum_{j=1}^{M} \frac{MV_{nzj}}{MV_{cj}} + \frac{1}{N} \sum_{k=1}^{N} \frac{MV_{nzk}}{MV_{ck}} \right). \qquad (1)$$

In this equation, α is the coefficient and $Bits_I$ and $Bits_P$ are the numbers of coded I-frame and P-frame bits. To normalize bits, $Bits_I$ and $Bits_P$ are divided by the maximum possible number of bits in a video frame Max_{bits}. L and S represent the number of I and P frames, and M and N denote the number of B and P frames, respectively. $\frac{MV_{nz}}{MV_c}$ represents the ratio of non-zero motion vectors MV_{nz} to the total number of counted motion vectors MV_c per picture.

Mean opinion score (MOS) with a 5-point absolute category rating (ACR) scale is calculated with a logarithmic relationship with CT and linear relationship with QP as follows:

$$MOS = \alpha + \beta \times (QP) + \gamma \times \ln(CT) \qquad (2)$$

The second model is ITU-T P.1204 model [21]. ITU-T standardized a bitstream-quality-estimation model for 2D-UHD video and this model was verified by using many subjective data. This model has two parts: parametric and machine learning. The parametric part takes the display and encoded resolution, framerate, and QP as input and can be used for estimating the quality with a 5-point ACR scale. In this part, quantization degradation D_{q}, upscaling degradation D_{u}, and temporal degradation D_{t} are calculated with QP ($quant$), display, encoded resolution, and framerate, respectively. The overall MOS of the parametric part $M_{\mathrm{parametric}}$ is calculated as follows:

$$D_{\mathrm{q}} = 100 - \max(\min(q_1 + q_2 \times \exp(q_3 \times quant + q_4)), 1), \qquad (3)$$

$$D_{\mathrm{u}} = \max(\min(x \times \log\left(y \times \ scaleFactor\right) 100), 0), \qquad (4)$$

$$D_{\mathrm{t}} = z \times \log(k \times \mathrm{framerate\_scale\_factor}), \qquad (5)$$

$$\mathrm{scaleFactor} = \max\left(\frac{\mathrm{display\_res}}{\mathrm{coding\_res}}, 1\right), \qquad (6)$$

$$\mathrm{framerate\_scale\_factor} = \frac{\mathrm{codingframerate}}{60}, \qquad (7)$$

$$D = D_{\mathrm{q}} + D_{\mathrm{u}} + D_{\mathrm{t}}, \qquad (8)$$

$$M_{\mathrm{randomforest}} = MOSfromR(100 - D), \qquad (9)$$

where q_1-q_4, x, y, z, k are coefficients.

Overall quality is calculated by using the weighted average of parametric part MOS and machine learning part MOS. In the machine learning part, $Residual$ estimated by the parametric part and subjective quality (i.e., residual quality) is calculated by random forest regression. The output MOS of the machine learning part $M_{\mathrm{randomforest}}$ and overall quality Q are calculated as follows:

$$M_{\text{randomforest}} = M_{\text{parametric}} + Residual, \tag{10}$$

$$Q = w_1 \times M_{\text{parametric}} + w_2 \times M_{\text{randomforest}}, \tag{11}$$

where coefficients are $w_1 = 0.5$ and $w_2 = 0.5$.

These two models are for estimating the quality of 2D video. Since these 2D video quality-estimation models do not take parameters of high- and low-resolution-tile video as input, the averages of all parameters of the high- and low-resolution tiles (i.e., the QP of equation (2) and $quant$ of equation (3) is the average of the two values: averages of all high-resolution tiles' QP and low-resolution tiles' QP) are used as the input of these models.

The third model is the Shaowei model [47], in which the quality of low-quality tiles and delay are considered. The overall quality Q is calculated as follows:

$$Q = Q_{\text{Hst}} \cdot Q_{\text{NQQT}}(\tau, q_l) \cdot Q_{\text{NQST}}(\tau, s_l), \tag{12}$$

where Q_{Hst} is the highest quality in the experiment and τ is delay. s_l is the low-resolution tile resolution, and q_l is quantization stepsize. The normalized quality of QP impact with respect to the delay is denoted as NQQT, and the normalized quality of resolution impact is denoted as NQST. If the delay is zero, users can always watch high-quality tiles, so NQQT and NQST are 1. Therefore, $Q_{\text{NQQT}}(\tau, q_l)$ and $Q_{\text{NQST}}(\tau, s_l)$ are calculated as follows:

$$Q_{\text{NQQT}}(\tau, q_l) = a(q_l) \cdot e^{-b(q_l)\cdot\tau} + (1 - a(q_l)), \tag{13}$$

$$Q_{\text{NQST}}(\tau, s_l) = a(s_l) \cdot e^{-b(s_l)\cdot\tau} + (1 - a(s_l)), \tag{14}$$

where $a(q_l)$, $b(q_l)$, $a(s_l)$, and $b(s_l)$ are coefficients. The separable response of the q_l- and s_l-impact on the perceptual quality are considered with respect to the delay. In this paper, the parameters of $a(q_l)$ and $b(q_l)$ are calculated as follows:

$$a(q_l) = \frac{k_{aq1}}{1 + k_{aq2} \cdot q_l^{k_{aq3}}}, \tag{15}$$

$$b(q_l) = \frac{k_{bq1}}{1 + k_{bq2} \cdot q_l^{k_{bq3}}}, \tag{16}$$

where k_{aq1}, k_{aq2}, k_{aq3}, k_{bq1}, k_{bq2}, and k_{bq3} are coefficients. The parameters of $a(s_l)$ and $b(s_l)$ are calculated as follows:

$$a(s_l) = k_{as1} \cdot e^{-k_{as2}\cdot s_l} + k_{as3}, \tag{17}$$

$$b(s_l) = k_{bs1} \cdot e^{-k_{bs2}\cdot s_l} + k_{bs3}. \tag{18}$$

3. Proposed Model

This section describes the proposed bitstream-quality-estimation model that uses the quality of high- and low-resolution tiles and delay to estimate the quality of tile-based VR video.

3.1 Structure of Proposed Model

To develop a bitstream-quality-estimation model for tile-based VR video streaming, both high- and low-resolution tiles' quality needs to be taken into account. Therefore, in calculating quality of tile-based VR videos, three points are important.

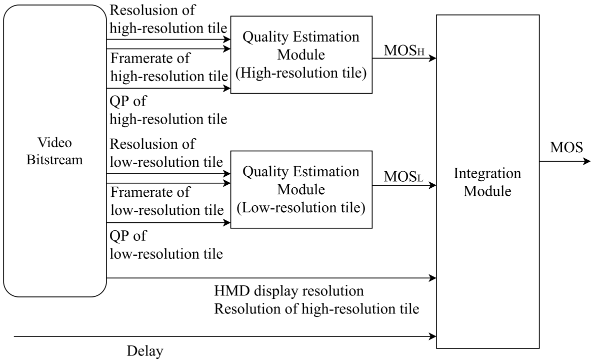

Video Bitstream
Resolution of high-resolution tile
Framerate of high-resolution tile
QP of high-resolution tile
Resolution of low-resolution tile
Framerate of low-resolution tile
QP of low-resolution tile
Quality Estimation Module (High-resolution tile)
Quality Estimation Module (Low-resolution tile)
MOSH
MOSL
Integration Module
MOS
HMD display resolution
Resolution of high-resolution tile
Delay

Fig. 2 Block diagram of proposed model.

- The quality of high-resolution tiles on the quality.
- The effect of resolution increase of low-resolution tiles on quality.
- The effect of delay (the display time of the low-resolution tiles) on the quality.

A block diagram of the proposed model is shown in Fig. 2. From the video bitstream, resolution, framerate, and QP of the high- and low-resolution tiles are extracted. In each quality-estimation part, high- and low-resolution tiles' quality are calculated, and the output quality of each tile is inputted into the integration module. In the integration module, overall MOS is calculated by high- and low-resolution tiles' quality. In this module, to take into account changing viewing directions, delay and resolution are inputted. Since delay is not constant, we estimate average delay calculated from the settings of the tile-based VR video, such as the chunk size of the tile-based VR video and the switching time of high-resolution tiles.

3.2 Quality-Estimation Module of High- and Low-Resolution Tiles

First, the quality of high- and low-resolution tiles $MOS_{(H,L)}$ is calculated. In our proposed model, the suffixes H and L represent the information of the high- and low-resolution tiles, respectively. To determine the maximum quality for each framerate and resolution, we introduce $X_{(H,L)}$, which is determined by framerate and resolution. The maximum MOS $X_{(H,L)}$ increases if the resolution or framerate increases. Lower QP makes video quality higher and MOS saturate at $X_{(H,L)}$, and higher QP makes lower quality and MOS saturate at 1. To express these characteristics, we use these equations to calculate the quality of high- and low-resolution tiles as follows:

$$X_{(H,L)} = \frac{4 \times \left(1 - \exp\left(-v_{3(H,L)} \times r\right)\right) \times s}{v_{2(H,L)} + s} + 1, \tag{19}$$

$$Y_{(H,L)} = \frac{s}{v_{4(H,L)}} + v_{5(H,L)} \times \log_{10}\left(v_{6(H,L)} \times r + 1\right), \tag{20}$$

$$MOS_{(H,L)} = X_{(H,L)} + \frac{1 - X_{(H,L)}}{1 + \left(\frac{QP_{(H,L)}}{Y_{(H,L)}}\right)^{v_{1(H,L)}}}. \tag{21}$$

In these equations, r represents the framerate of the video, s represents the resolution of the tile, which is the number of pixels per frame (i.e., width × height), and $v_1 - v_6$ are coefficients of the model. QP is the average QP of P and B frames in block units on overall frames. The parameter $Y_{(H,L)}$ represents the inflection point of the graph of QP versus MOS.

3.3 Integration Module of High- and Low-Resolution Tiles Quality

As described previously, when users change viewing directions while watching a tile-based VR video, users watch a low-resolution tile. Since the low-resolution tiles are enlarged, users perceive quality degradation of the low-resolution tiles. The ratio of users viewing low-resolution tile time to users viewing high-resolution tile time is affected by the percentage of the HMD display area occupied by high-resolution tiles and delay. To take these perspectives into account, we introduce the parameter ocr, which represents the percentage of the HMD display occupied by high-resolution tiles. We also introduce the parameter $delay$, which represents delay in switching from high-resolution tiles to low-resolution tiles. The overall quality of tile-based VR video is calculated as follows:

$$\text{MOS} = a \times MOS_H + (1 - a) \times MOS_L, \tag{22}$$

where the coefficient a is the ratio that determines the quality of the high- and low-resolution tiles and a is expressed as follows:

$$a = v_7 \times delay^{-v_8} + v_9 \times ocr, \tag{23}$$

$$ocr = \min\left(\frac{Res_H}{Res_{HMD}}, 1\right), \tag{24}$$

where $v_7 - v_9$ are coefficients. Res_H is the resolution of high-resolution tiles, and Res_{HMD} is the resolution of the HMD display. In this experiment, the resolutions of high- and low-resolution tiles are the same. Equation (23) shows shorter $delay$ and larger ocr make the parameter a larger, so the ratio of high-resolution tiles becomes larger and the percentage of high-resolution tiles in the overall MOS becomes larger.

4. Subjective Experiment

This section describes subjective experiments. We conducted subjective experiments by varying the quality degradations of high- and low-resolution tiles and delay. To check the accuracy of the proposed model, two experiments were carried out: one is used as training data and the other as test data. The procedure and settings of these experiments are shown in this section.

Church Soccer
Dance Drive
Fall Flower

Fig. 3 Experiment 1 SRCs.

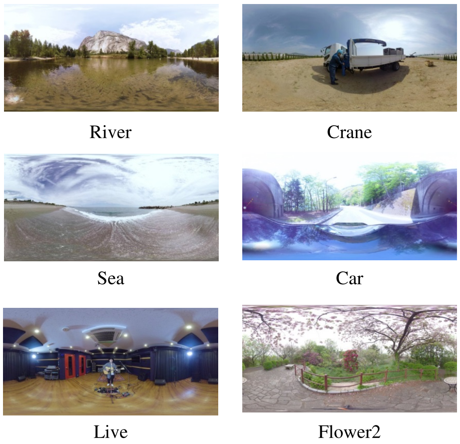

River Crane
Sea Car
Live Flower2

Fig. 4 Experiment 2 SRCs.

4.1 Source Reference Circuits

Twelve source videos, also called source reference circuits (SRCs), are used in two experiments. The duration of SRC was 20 seconds. Six SRCs are assigned to each of the two experiments. Figures 3 and 4 show the SRCs of experiments 1 and 2, and the details of videos are shown in Table 1. The resolution of source videos was 7680 × 3840/30 fps (chroma sampling: 4:2:0). SRCs were characterized in terms of their spatial information (SI) and temporal information (TI) [48]. In terms of calculating 360° equirectangular VR video SI and TI, the distortion and warping of the video pole have to be considered. To determine the average feature of SRCs, the average SI and TI of all frames in SRCs are calculated in the spherical domain [49], [50]. Figures 5 and 6 show the spherical SI and TI in both experiments, and the SRCs are found to have different motions and edge features. For watching VR video naturally, the stereo channel audio was used in this experiment.

Table 1 Details of SRCs used in experiments.

Experiment	SRC	Type of contents
1	church	In a church, little motion and many colors
	soccer	Penalty kick practice, a lot of motion
	dance	Three women dance orbiting around the camera, a lot of motion
	drive	Video from inside a car driving on a highway
	waterfall	A scene of a waterfall, and the splash has detailed textures
	flower	A scene of a flower garden, little motion and detailed textures
2	river	A scene of a river, with the water surface vibrating
	crane	A man is operating a crane, it is moving slightly
	sea	A scene of the sea, the wave have detailed textures
	car	A video from inside a car in a forest
	live	A man is playing guitar in a studio, with few motions
	flower2	A scene of a flower park, little motion and detailed textures

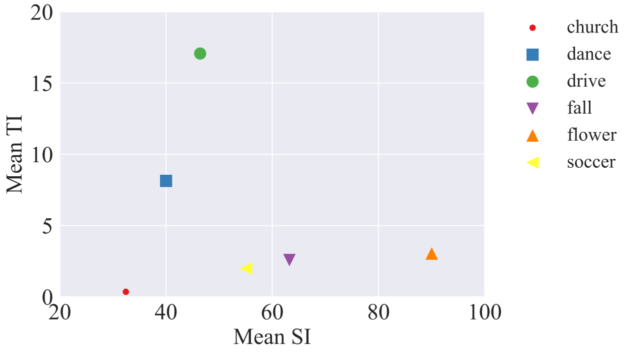


Fig. 5 SI and TI of SRCs in experiment 1.

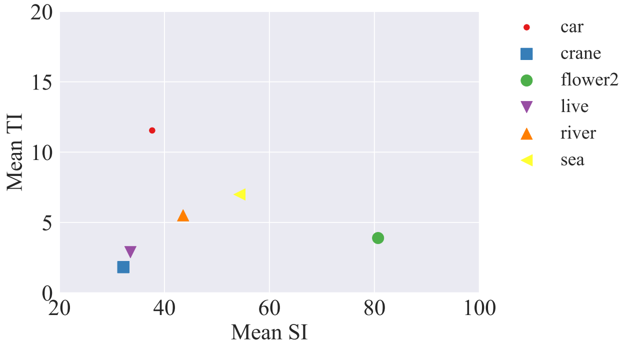


Fig. 6 SI and TI of SRCs in experiment 2.

4.2 Experimental Conditions

To verify the effect of high- and low-resolution tiles parameters and delay on subjective quality, resolution, bitrate, and delay are varied as experimental conditions, also called hypothesis reference circuits (HRCs). To encode from 360° videos to tile-based VR videos, a tiled-based encoding [5] was used, as shown in Fig. 7. SRCs were encoded by FFmpeg encoder v3.0, and all tiles were encoded by H.265/high-efficiency video coding (HEVC) (Main Profile/Level 5, GOP: M = 3, N = 15). The segment size was set to 0.5 seconds, and the chroma sampling was 4:2:0.

High and low bitrates, delay, resolution of the experiments are shown in Tables 2 and 3. These tables show the high/low bitrate pairs, delay, and resolution. Several bitrate

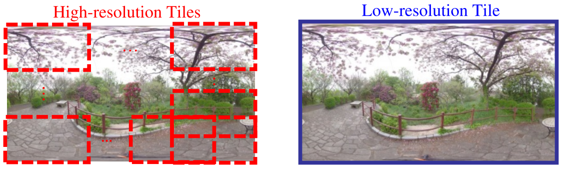


Fig. 7 Tile-based VR video.

Table 2 HRC in Experiment 1.

High/Low bitrate (Mbps)	Delay	Resolution
40/20	1	960
2/1	3	960
10/2.5	5	960
40/20	10	960
40/20	1	1280
2/1, 10/10	3	1280
2/2, 40/20	5	1280
40/10, 2/1	10	1280
40/40, 10/10, 2/2, 40/20, 10/5, 40/10, 10/2.5, 2/1, 2/0.5	1	1920
2/0.5	3	1920
10/5, 40/10	5	1920
10/2.5, 2/2	10	1920
40/20	1	3840
40/10	3	3840
40/40	5	3840
10/5	10	3840

Table 3 HRC in Experiment 2.

High/Low bitrate (Mbps)	Delay	Resolution
6/2	1	960
30/30, 6/4	2	960
40/30	4	960
20/5, 4/0.5	8	960
8/4	2	1280
4/2	4	1280
20/20, 20/8	6	1280
40/10, 30/20	8	1280
6/3, 2/0.5	10	1280
20/10, 10/6, 10/2, 4/1	1	1920
4/2	4	1920
8/8, 4/0.5	6	1920
10/8, 6/2	8	1920
30/30, 30/10	10	1920
20/10, 8/2, 2/1	1	2560
30/10, 4/4	2	2560
40/20, 8/6	4	2560
4/1	6	2560
8/6	10	2560
40/5, 30/5, 8/8, 2/1	1	3840
2/0.5	2	3840
20/15,10/10,2/2	6	3840
10/4	10	3840

pairs are listed in the several rows of the bitrate column. This means that each bitrate pair is used to encode at the listed delay and resolution in the row.

High-resolution tiles are encoded at 9 bitrate levels between 2 and 40 Mbps, and low-resolution tiles are encoded at 10 bitrate levels between 0.5 and 40 Mbps. Resolution conditions are as follows: 3840×3840, 2560×2560, 1920×1920, 1280×1280, and 960×960, where 2560×2560 is used in experiment 2 only. The high-resolution tiles are not down-converted and display the original resolution. The framerate of the tile-based VR video was 30 fps. The high-resolution tiles were divided into 5×12 tiles (except for 3840×3840), and 1×12 tiles (3840×3840)). All high-resolution tiles are overlapped by blocks, as shown in Fig. 7 to avoid quality degradation due to small changes in the user's viewing direction. All high-resolution tiles are arranged at equal intervals over the entire horizontal or vertical space, and the degree of overlap depends on the resolution. The number of pixel of horizontal direction overlap is $(\text{tilesize\_pixel}\times 12-7680)/12$, where tilesize_pixel represents resolution i.e., 3840, 2560, 1920, 1280 or 960 in this experiment. The number of pixels of vertical direction overlap (except for 3840×3840) is $(\text{tilesize\_pixel}\times 4-3840)/5$ because the tiles on both ends are connected, but the tiles on the top and bottom rows are not connected. The low-resolution tile consists of one tile and is encoded at the same resolution as high-resolution tiles. Then the low-resolution tile was enlarged to the original resolution (i.e., 7680×3840) at the HMD to display the 360° videos, so the low-resolution tile was degraded by upscaling.

Audio was encoded by AAC-LC. Bitrate and sampling rate were 128 kbps and 48 kHz, respectively.

Delay conditions are 1 to 10 seconds (1, 3, 5, and 10 seconds in experiment 1 and 1, 2, 4, 6, 8, and 10 seconds in experiment 2). When users change their viewing directions, HMD sends it to the server and the server distributes the new high-resolution tile. Delay is inserted when the server sends the new high-resolution tile. The minimum delay is 1 second because it takes about 1 second to reflect the viewing direction.

Each HRC is assigned 1 or 2 SRCs and in total, 96 (Experiment 1) and 99 (Experiment 2) videos were generated, which are called processed video sequences (PVSs). To determine the relationship of MOS between experiments 1 and 2, 9 anchor videos are used in both experiments. The anchor conditions and SRCs are shown in Table 4.

Table 4 Anchor PVSs.

High/Low bitrate (Mbps)	Delay	Resolution	SRC
2/2	10	1280	church
2/2	1	1920	soccer
10/10	1	1920	waterfall
40/40	1	1920	church
2/0.5	3	1920	flower
10/2.5	10	1920	dance
40/20	10	1920	drive
40/10	3	3840	waterfall
10/5	10	3840	soccer

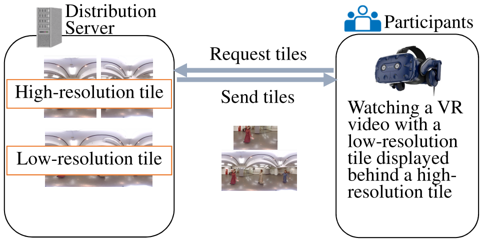


Fig. 8 Distribution system.

4.3 Subjective Experiment Environment and Procedures

In this section, we explain the method and environment. The experimental system is shown in Fig. 8, where the distribution server sends high- and low-resolution tiles. During the experiment, participants sat on a revolving chair, were instructed that they could change viewing directions, and watched the 360° videos freely by rotating the chair. Participants used Vive Pro, which is a VR headset with two rectangular glasses-like screens (1440×1600 each), to watch tile-based VR videos. For watching VR video naturally, the stereo channel audio was used, where a comfortable audio listening level was selected by them and the audio volume was not changed during the experiment. Therefore, the audio volume is the same between the PVSs that are the same SRCs.

Before conducting these subjective experiments, participants read the instructions to understand the procedure and the objective of the subjective experiment. In addition, participants were told that video quality needs to be evaluated by not taking into account audio quality. After taking visual acuity and color vision tests, participants took a practice session in which they were told how to wear the HMD and the method to evaluate PVSs. After this explanation, participants watched 20-second test videos 6 times. Between videos, participants scored the video with a 5-point ACR method using a controller displayed on the HMD. Three seconds after scoring, the next PVS started.

After participants learned the procedure and method to evaluate videos in the practice session, the test session started. Participants watched 20-second PVSs 6 times in 1 session that lasted about 3.5 minutes. Three seconds after participants had scored, the next PVS started. There were 2-minute breaks between sessions, and 5- to 10-minute breaks every 3 to 4 sessions.

The total experiment lasted about 150 minutes, including visual acuity and color vision tests, instruction, tests, and breaks. The presentation order of the PVSs was randomized.

4.4 Participants

In each experiment, 36 participants passed visual acuity and color vision tests. In experiment 1, participants were 18 males and 18 females ranging from 18 to 31 years old, with an average age of 21.0. In experiment 2, participants were 18 males and 18 females ranging from 18 to 25 years old,

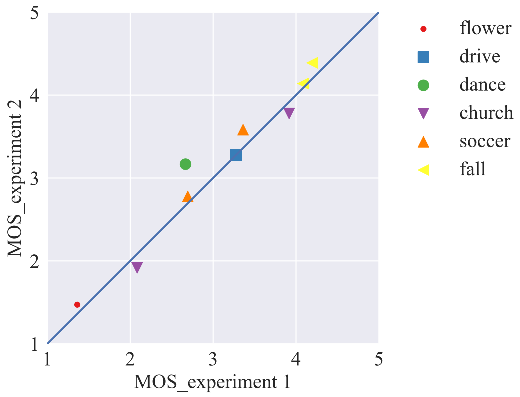


Fig. 9 Anchor PVSs MOS.

Table 5 Estimation results: training data is experiment 1 and test data is experiment 2.

	Training (Experiment 1)			Test (Experiment 2)		
Models	RMSE	PCC	SROCC	RMSE	PCC	SROCC
Model 1	0.401	0.899	0.894	0.536	0.760	0.782
Model 2	0.349	0.924	0.927	0.456	0.835	0.848
Model 3	0.282	0.951	0.948	0.443	0.843	0.840
Model 4	0.241	0.965	0.964	0.350	0.905	0.906

Table 6 Estimation results: training data is experiment 2 and test data is experiment 1.

	Training (Experiment 2)			Test (Experiment 1)		
Models	RMSE	PCC	SROCC	RMSE	PCC	SROCC
Model 1	0.472	0.819	0.843	0.462	0.868	0.897
Model 2	0.391	0.880	0.891	0.406	0.920	0.919
Model 3	0.355	0.902	0.904	0.341	0.929	0.932
Model 4	0.289	0.936	0.939	0.316	0.940	0.944

with an average age of 20.9.

5. Results

This section describes the experimental results, and the accuracy of the proposed quality-estimation model is compared with those of other video estimation models explained in Sect. 2.

5.1 Anchor PVS Analysis

First, to make sure that the bias between the two experiments was small, we confirmed anchor PVSs' MOSs. Anchor PVSs' MOSs between experiments 1 and 2 are shown in Fig. 9. In Fig. 9, anchor PVSs' MOSs are almost lined up on a 45-degree line and have a strong relationship between the two experiments. Therefore, we use the original MOS when we cross-validate between experiments 1 and 2.

5.2 Quality-Estimation Accuracy

In Sect. 2, the effects of low-resolution tile and delay are stated. To evaluate the validity of the proposed model with the low-resolution tile parameters and delay, the quality-estimation accuracies of the models with and without these parameters are calculated. We compared four models by calculating their accuracies:

Model 1 With high-resolution tile parameters
Model 2 With mean of high- and low-resolution tile parameters
Model 3 With high- and low-resolution tile parameters with weighed coefficients
Model 4 With high- and low-resolution tile parameters and delay with weighed coefficients

All models used resolution, framerate, and QP as input, and we optimized the coefficients of each model using the subjective quality data, respectively. Model 1 uses high-resolution tiles' QP but not low-resolution tiles' QP and optimizes $v_{1(H)} - v_{6(H)}$. Model 2 uses averaged high- and low-resolution tiles' QP, i.e., uses $\frac{(QP_H+QP_L)}{2}$ and optimizes $v_{1(H)} - v_{6(H)}$. Model 3 uses both high- and low-resolution tiles' QP but not delay, so $v_8 = 0$, and it optimizes $v_{1(H,L)} - v_{6(H,L)}, v_7, v_9$. Model 4 uses high- and low-resolution tiles' QP and delay and optimizes all coefficients, $v_{1(H,L)} - v_{6(H,L)}, v_7 - v_9$. These coefficients are optimized using the least-squares method by curve fitting with Python Version 3.4.9. In each experiment, we optimized the coefficients as training data. Also, to determine the effect of training data, we conducted cross-validation with each experiment. Tables 5 and 6 show the relationship between estimated MOS and subjective MOS for each tile-based VR video using optimized coefficients. The left side of Table 5 shows the results of Experiment 1 as training, and the right side shows the results of using Experiment 1 as the training data and validating it with Experiment 2. Similarly, the left side of Table 6 shows the results of Experiment 2 as training, and the right side shows the results of using Experiment 2 as the training data and validating it with Experiment 1. Root mean square error (RMSE), Pearson correlation coefficient (PCC) and Spearman's rank-order correlation coefficient (SROCC) between subjective MOS and estimated MOS are shown. As shown in Tables 5 and 6, quality-estimation accuracy was improved by applying low-resolution tiles' QP and delay to the proposed model.

5.3 Comparison with Other Comparative Models

To evaluate the performance of the proposed model, we compared the results of the proposed model and other quality-estimation models [21], [39], [47] explained in Sect. 2.3. To apply these 2D quality-estimation models [21], [39] to a tile-based VR quality-estimation model, resolution, framerate, and the average of high- and low-resolution tiles' QP are used as input of the models. In other words, the input QP of these models is the average of two parameters: the averages of 60 high-resolution tiles' QP and 60 low-resolution tiles' QP. To apply the Shaowei model as a comparative model, we convert the quantization parameter into quantization stepsize as $q = 2^{\frac{QP-4}{6}}$, and Q_{Hst} to our experiment's highest quality. Tables 7 and 8 compare the test data results between the proposed model and other models. Each model is cross-validated, i.e., training data is experiment 1 and test data is experiment 2, and training data is experiment 2 and test data is experiment 1. Performance values for all models are given in these tables. According to these tables, the RMSEs of the proposed model are 0.350 and 0.316, which are the best scores of all models. These results show that the proposed model can estimate the tile-based VR video accurately. By introducing low-resolution tile weighting and delay, the proposed method can estimate quality with higher accuracy than the P.1204 and Anegekuh models. Also, the proposed model is found to have the better quality estimation accuracy in terms of RMSE and PCC than the model that considers low-resolution tile and delay [47]. Since the orders of estimated quality and subjective quality of the proposed model are different especially in SRC of dance and flower, its SROCC is slightly worse than the that of Shaowei model. However, the difference in SROCC between the proposed and Shaowei models was small. Therefore, the proposed model can be concluded to have a good quality estimation accuracy because it has better RMSE and PCC than the Shaowei model.

5.4 Discussion of Estimation Accuracy for SRCs

To check the accuracy of estimated quality for each SRCs, we compared the actual and estimated MOSs for each SRCs. Figures 10 and 11 show the estimated MOS and subjective MOS for test datasets using the proposed model. Figure 10 shows the test data results of Experiment 2 with Experiment 1 as training, and Fig. 11 shows the test data results of Experiment 1 with Experiment 2 as training. These figures are color-coded by SRCs. As indicated in Figs. 10 and 11, the tendency of estimation accuracy is different from that of SRCs. To clarify the video effect of video prosperities, such as motion and complexity of the video, we checked the relationship of SI, TI and difference between the estimated MOS and the subjective MOS (deltaMOS), as shown in Fig. 12. DeltaMOS represents estimated MOS minus subjective MOS, so larger deltaMOS indicates that the proposed model overestimates MOS. The vertical axis in Fig. 12 is the averaged deltaMOS over PVSs of the same SRC. These figures show that the estimated MOS tends to be higher when the TI is large. In these high TI SRCs, quality is degraded due to movement that is not fully captured by QP. On the other hand, the estimated MOS of the crane is lower than the subjective MOS. This SRC has low SI and TI, and colors are more monotonous than in other SRCs, and participants had slight difficulty perceiving the tile-based VR video degradation. In this experiment, the effect of SI on quality differs from experiment 1 and 2, so the effect could not be determined. The effect of SI on quality needs to be clarified in more detail by conducting experiments using a large number of SRCs. In this study, pixel information is not used, and TI of the tile-based VR video cannot be used. However, it is considered that the accuracy of the quality estimation can be improved by using detailed information such as motion vectors from the bitstream to estimate the degree of motion of the tile-based VR video.

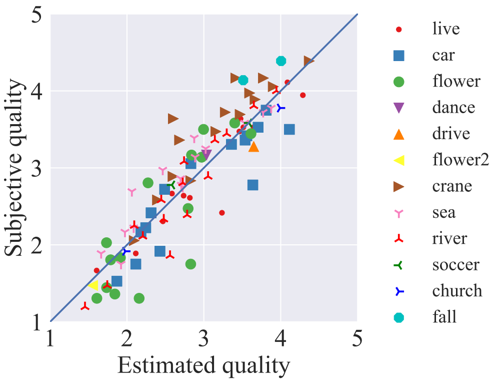


Fig. 10 Estimation results: training data is experiment 1 and test data is experiment 2.

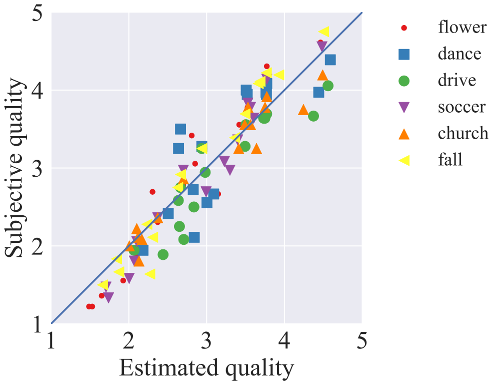


Fig. 11 Estimation results: training data is experiment 2 and test data is experiment 1.

Table 7 Comparison between proposed model and other models: training data is experiment 1 and test data is experiment 2.

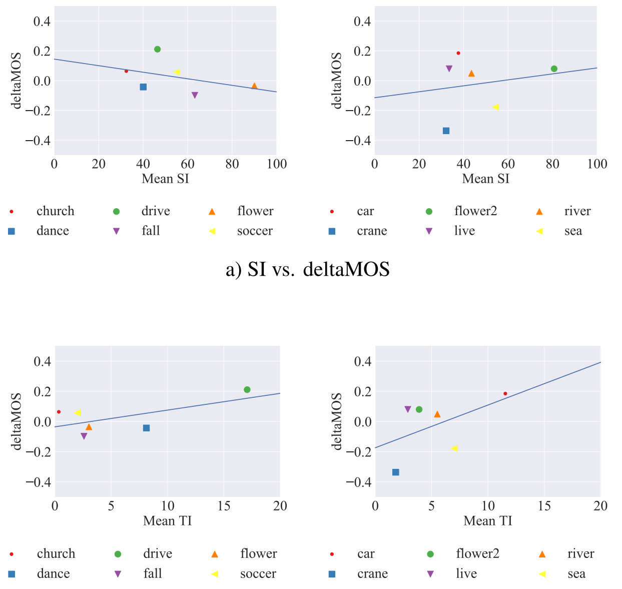
Model	RMSE	PCC	SROCC
Proposed model	0.350	0.897	0.884
P.1204 model	0.455	0.836	0.837
Anegekuh model	0.541	0.769	0.682
Shaowei model	0.397	0.889	0.891

Table 8 Comparison between proposed model and other models: training data is experiment 2 and test data is experiment 1.

Model	RMSE	PCC	SROCC
Proposed model	0.316	0.939	0.938
P.1204 model	0.466	0.861	0.872
Anegekuh model	0.536	0.859	0.707
Shaowei model	0.333	0.933	0.940

6. Usage of Proposed Model

This section describes how to use the proposed model. To select a suitable bitrate for tile-based VR video streaming,



a) SI vs. deltaMOS

b) TI vs. deltaMOS

Fig. 12 SI and TI vs. deltaMOS.

the bitrate and quality need to be checked at the same time. If low-quality content is found, the quality is improved by increasing the bitrate. Concretely, a content is encoded at several bitrates and the quality is calculated by the proposed model. Then, the relationship between bitrate and quality is obtained. By using this relationship, a suitable bitrate is found on the basis of the service providers' desired quality.

7. Conclusion

To monitor tile-based virtual reality (VR) video quality at the head end, we proposed a bitstream-quality-estimation model, which uses high- and low-resolution tiles' quantization parameter (QP), resolution, and delay. To validate our proposed model, two subjective experiments were performed. First we found that the proposed model accuracy was improved by introducing the parameters of high- and low-resolution tiles, resolution, and delay. Then the quality-estimation accuracy was compared between the proposed model and other quality-estimation models used in quality estimation. The coefficients of each model were optimized, and the results were cross-validated between two experiments. The results revealed that the proposed model can estimate tile-based VR video quality more accurately than other video quality-estimate models. Quality-estimation accuracy of the proposed model is improved by using low-resolution tiles' QP and delay. Also the results suggest that the proposed model can estimate the quality with high accuracy between the training and test datasets.

The purpose of the proposed model is to estimate tile-based VR video coding quality. However, if this model is to be used for monitoring session quality of tile-based VR video, something like the long-time model on P.1203 is needed. In future work, to improve the quality-estimation accuracy, we will introduce detailed information such as motion vectors from the bitstream. To check the effect of SI on the tile-based VR video quality, experiments with many SRCs have to be conducted. To investigate how audio quality affects tile-based VR video quality, audio quality will be changed in the future experiments.

References

[1] K.K. Sreedhar, A. Aminlou, M.M. Hannuksela, and M. Gabbouj, "Viewport-adaptive encoding and streaming of 360-degree video for virtual reality applications," 2016 IEEE International Symposium on Multimedia (ISM), pp.583–586, Dec. 2016.

[2] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi, C. Chiang, Y. Wang, P. Wilkins, J. Bankoski, L. Trudeau, N. Egge, J. Valin, T. Davies, S. Midtskogen, A. Norkin, and P. de Rivaz, "An overview of core coding tools in the AV1 video codec," 2018 Picture Coding Symposium (PCS), pp.41–45, June 2018.

[3] 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, Dec. 2012.

[4] MPEG, "Omnidirectional media application format," WD on ISO/IEC 23000-20 Omnidirectional Media Application Format.

[5] H. Kimata, D. Ochi, A. Kameda, H. Noto, K. Fukazawa, and A. Kojima, "Mobile and multi-device interactive panorama video distribution system," The 1st IEEE Global Conference on Consumer Electronics 2012, pp.574–578, Oct. 2012.

[6] D. Ochi, Y. Kunita, A. Kameda, A. Kojima, and S. Iwaki, "Live streaming system for omnidirectional video," 2015 IEEE Virtual Reality (VR), pp.349–350, March 2015.

[7] L. Zhao, Y. Cui, Z. Liu, Y. Zhang, and S. Yang, "Adaptive streaming of 360 videos with perfect, imperfect, and unknown FoV viewing probabilities in wireless networks," IEEE Trans. Image Process., vol.30, pp.7744–7759, July 2021.

[8] C. Guo, Y. Cui, and Z. Liu, "Optimal multicast of tiled 360 VR video," IEEE Wireless Commun. Lett., vol.8, no.1, pp.145–148, Aug. 2019.

[9] J. van der Hooft, M. Torres Vega, S. Petrangeli, T. Wauters, and F. De Turck, "Quality assessment for adaptive virtual reality video streaming: A probabilistic approach on the user's gaze," 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), pp.19–24, Feb. 2019.

[10] D.V. Nguyen, H.T.T. Tran, and T.C. Thang, "An evaluation of tile selection methods for viewport-adaptive streaming of 360-degree video," ACM Trans. Multimedia Comput. Commun. Appl., vol.16, no.1, pp.1–24, March 2020.

[11] C. Cortes, P. Perez, J. Gutierrez, and N. Garcia, "Influence of video delay on quality, presence, and sickness in viewport adaptive immersive streaming," 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX), pp.1–4, May 2020.

[12] M. Orduna, C. Diaz, L. Munoz, P. Perez, I. Benito, and N. Garcia, "Video multimethod assessment fusion (VMAF) on 360 VR contents," IEEE Trans. Consum. Electron., vol.66, no.1, pp.22–31, Jan. 2020.

[13] W. Sun, K. Gu, G. Zhai, S. Ma, W. Lin, and P. Le Calle, "CVIQD: Subjective quality evaluation of compressed virtual reality images," 2017 IEEE International Conference on Image Processing (ICIP), pp.3450–3454, Sept. 2017.

[14] A. Singla, W. Robitza, and A. Raake, "Comparison of subjective quality evaluation methods for omnidirectional videos with DSIS and modified ACR," Electronic Imaging, vol.30, no.14, pp.1–6, 2018.

[15] A. Raake, S. Borer, S.M. Satti, J. Gustafsson, R.R.R. Rao, S. Medagli, P. List, S. Göring, D. Lindero, W. Robitza, G. Heikkilä, S. Broom, C. Schmidmer, B. Feiten, U. Wüstenhagen, T. Wittmann, M. Obermann, and R. Bitto, "Multi-model standard for bitstream-, pixel-

based and hybrid video quality assessment of UHD/4K: ITU-T P.1204," IEEE Access, vol.8, pp.193020–193049, Oct. 2020.

[16] "Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full and reduced reference pixel information," Recommendation ITU-T P.1204.4, Jan. 2020.

[17] "Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference," Recommendation ITU-T J.246, Aug. 2008.

[18] "Final report from the video quality experts group on the validation of objective models of video quality assessment," VQEG, Sept. 2003.

[19] "Final report from the video quality experts group on the validation of objective models of video quality assessment, Phase 2," VQEG, Aug. 2000.

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

[21] "Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full bitstream information," Recommendation ITU-T P.1204, Jan. 2020.

[22] "Performance monitoring points for IPTV," Recommendation ITU-T G.1081, Oct. 2008.

[23] M.A. Layek, N.Q. Thai, M.A. Hossain, N.T. Thu, L.P. Tuyen, A. Talukder, T. Chung, and E. Huh, "Performance analysis of H.264, H.265, VP9 and AV1 video encoders," 2017 19th Asia-Pacific Network Operations and Management Symposium (APNOMS), pp.322–325, Sept. 2017.

[24] Q. Huangyuan, L. Song, Z. Luo, X. Wang, and Y. Zhao, "Performance evaluation of H.265/MPEG-HEVC encoders for 4K video sequences," Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, pp.1–8, Feb. 2014.

[25] H. Duan, G. Zhai, X. Yang, D. Li, and W. Zhu, "IVQAD 2017: An immersive video quality assessment database," 2017 International Conference on Systems, Signals and Image Processing (IWSSIP), pp.1–5, May 2017.

[26] H.T. Tran, N.P. Ngoc, C.T. Pham, Y.J. Jung, and T.C. Thang, "A subjective study on QoE of 360 video for VR communication," 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), pp.1–6, Oct. 2017.

[27] M. Yang, W. Zou, and F. Yang, "A subjective perceptual quality evaluation for emerging cloud VR services," 2019 IEEE 5th International Conference on Computer and Communications (ICCC), pp.1226–1230, Dec. 2019.

[28] Y. Han, C. Yu, D. Li, J. Zhang, and Y. Lai, "Accuracy analysis on 360° virtual reality video quality assessment methods," 2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC), pp.414–419, Dec. 2020.

[29] Z. Fei, F. Wang, J. Wang, and X. Xie, "QoE evaluation methods for 360-degree VR video transmission," IEEE J. Sel. Top. Signal Process., vol.14, no.1, pp.78–88, Nov. 2020.

[30] A. Singla, S. Göring, A. Raake, B. Meixner, R. Koenen, and T. Buchholz, "Subjective quality evaluation of tile-based streaming for omnidirectional videos," Proc. 10th ACM Multimedia Systems Conference, MMSys'19, pp.232–242, Association for Computing Machinery, June 2019.

[31] C. Ozcinar, J. Cabrera, and A. Smolic, "Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality," IEEE Trans. Emerg. Sel. Topics Circuits Syst., vol.9, no.1, pp.217–230, Jan. 2019.

[32] Z.L. Xiaolan Jiang, Y.-H. Chiang, and Y. Ji, "Improving QoE of viewport adaptive 360-degree video streaming with machine learning," IEICE Technical Report, pp.49–54, May 2018.

[33] J. van der Hooft, M.T. Vega, S. Petrangeli, T. Wauters, and F. De Turck, "Quality assessment for adaptive virtual reality video streaming: A probabilistic approach on the users gaze," 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), pp.19–24, Feb. 2019.

[34] Y. Han, Y. Ma, Y. Liao, and G. Muntean, "QoE oriented adaptive streaming method for 360° virtual reality videos," 2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp.1655–1659, Aug. 2019.

[35] Y.S. de la Fuente, G.S. Bhullar, R. Skupin, C. Hellge, and T. Schierl, "Delay impact on MPEG OMAF's tile-based viewport-dependent 360 video streaming," IEEE Trans. Emerg. Sel. Topics Circuits Syst., vol.9, no.1, pp.18–28, Feb. 2019.

[36] M. Takagi, H. Fujii, A. Shimizu, Y. Urata, and K. Yamagishi, "Subjective video quality estimation to determine optimal spatio-temporal resolution," 2013 Picture Coding Symposium (PCS), pp.422–425, Dec. 2013.

[37] K. Izumi, K. Kawamura, T. Yoshino, and S. Naito, "No reference video quality assessment based on parametric analysis of HEVC bitstream," 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp.49–50, Sept. 2014.

[38] C. Keimel, M. Klimpke, J. Habigt, and K. Diepold, "No-reference video quality metric for HDTV based on H. 264/AVC bitstream features," 2011 18th IEEE International Conference on Image Processing, pp.3325–3328, Sept. 2011.

[39] L. Anegekuh, L. Sun, E. Jammeh, I.H. Mkwawa, and E. Ifeachor, "Content-based video quality prediction for HEVC encoded videos streamed over packet networks," IEEE Trans. Multimedia, vol.17, no.8, pp.1323–1334, Aug. 2015.

[40] B. Lee and M. Kim, "No-reference PSNR estimation for HEVC encoded video," IEEE Trans. Broadcast., vol.59, no.1, pp.20–27, March 2013.

[41] T. Brandão and M.P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," IEEE Trans. Circuits Syst. Video Technol., vol.20, no.11, pp.1437–1447, Dec. 2010.

[42] S. Fremerey, S. Goring, R. Rao, R. Huang, and A. Raake, "Subjective test dataset and meta-data-based models for 360° streaming video quality," 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), pp.1–6, Sept. 2020.

[43] M.S. Anwar, J. Wang, W. Khan, A. Ullah, S. Ahmad, and Z. Fei, "Subjective QoE of 360-degree virtual reality videos and machine learning predictions," IEEE Access, vol.8, pp.148084–148099, Aug. 2020.

[44] H.G. Kim, H. Lim, and Y.M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," IEEE Trans. Circuits Syst. Video Technol., vol.30, no.4, pp.917–928, Feb. 2020.

[45] J. Li, R. Feng, Z. Liu, W. Sun, and Q. Li, "Modeling QoE of virtual reality video transmission over wireless networks," 2018 IEEE Global Communications Conference (GLOBECOM), pp.1–7, Dec. 2018.

[46] R. Schatz, A. Zabrovskiy, and C. Timmerer, "Tile-based streaming of 8K omnidirectional video: Subjective and objective QoE evaluation," 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), pp.1–6, June 2019.

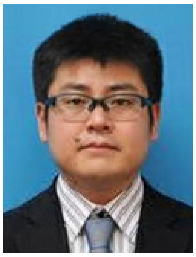
[47] S. Xie, Y. Xu, Q. Shen, Z. Ma, and W. Zhang, "Modeling the perceptual quality of viewport adaptive omnidirectional video streaming," IEEE Trans. Circuits Syst. Video Technol., vol.30, no.9, pp.3029–3042, Aug. 2020.

[48] "Subjective video quality assessment methods for multimedia applications," Recommendation ITU-T P.910, April 2008.

[49] Youncy-Hu, "Spherical SI/TI," https://github.com/Youncy-Hu/Spherical-SI-TI

[50] X. Liu, P. An, C. Meng, C. Yang, and X. Huang, "Multiscale WS-SSIM for panoramic video quality assessment," Optoelectronic Imaging and Multimedia Technology VII, Q. Dai, T. Shimura, and Z. Zheng, eds., pp.96–101, International Society for Optics and Photonics, Oct. 2020.

Masanori Koike received his B.E. degree in mathematical engineering and information physics in 2015 and M.S. degree in information science and technology in 2017 from the University of Tokyo, Japan. Since joining NTT Network Technology Laboratories in 2017, he has been engaged in research on quality of experience for VR video distribution.



Yuichiro Urata received his B.E. and M.E. degrees in engineering from University of Electro-Communications, Tokyo, Japan in 2009 and 2011. Since 2011, he has worked for NTT Network Technology Laboratories in Tokyo, where his work has focused on the quality of experience for videos.

Kazuhisa Yamagishi received his B.E. degree in electrical engineering from the Tokyo University of Science, Chiba, Japan, in 2001, and his M.E. and Ph.D. degrees in electronics, information, and communication engineering from Waseda University, Tokyo, Japan, in 2003 and 2013. He joined NTT Laboratories, Tokyo, Japan, in 2003. From 2010 to 2011, he was a visiting researcher with Arizona State University, Tempe, AZ, USA. His research interests include the development of objective quality-estimation models for multimedia telecommunications. He was a recipient of the Young Investigators' Award (IEICE) in Japan in 2007 and the Telecommunication Advancement Foundation Award in Japan, in 2008, the ITU-AJ Encouragement Awards in Japan, in 2017, and the TTC Award for Distinguished Service in Japan, in 2018.