IEICE TRANS. INF. & SYST., VOL.E101–D, NO.1 JANUARY 2018

PAPER *Special Section on Enriched Multimedia — Potential and Possibility of Multimedia Contents for the Future —*

A Study on Quality Metrics for 360 Video Communications

Huyen T. T. TRAN[†a)], **Cuong T. PHAM**[††], **Nam PHAM NGOC**[††], ***Nonmembers***, **Anh T. PHAM**[†], ***and*** **Truong Cong THANG**[†], ***Members***

SUMMARY 360 videos have recently become a popular virtual reality content type. However, a good quality metric for 360 videos is still an open issue. In this work, our goal is to identify appropriate objective quality metrics for 360 video communications. Especially, fourteen objective quality measures at different processing phases are considered. Also, a subjective test is conducted in this study. The relationship between objective quality and subjective quality is investigated. It is found that most of the PSNR-related quality measures are well correlated with subjective quality. However, for evaluating video quality across different contents, a content-based quality metric is needed.

key words: *360 video, objective quality, subjective quality, video processing*

1. Introduction

360-degree videos (or 360 videos for short) have become a popular virtual reality (VR) content type on video streaming platforms. While there are a lot of previous studies on traditional videos [1]–[3], the research on 360 videos is still very limited. Most of 360 video related studies focus on investigating some aspects of video quality [4]–[6] in VR environment such as the presence, usability, and cybersickness. Some recent studies consider quality optimization for 360 video delivery [7], [8]. However, a good quality metric for 360 videos is still an open issue.

Essentially, a spherical image of a 360 video needs to be converted to 2D plane so that it can be encoded by existing coding formats. This is supported by different projection types (e.g. Equi-rectangular projection (ERP) and Cube Map (CMP) projection). Obviously, such 2D images could be used as inputs for 360 video quality evaluation. In this way, two quality metrics, PSNR and Multiscale Structural similarity (MS-SSIM), are used to evaluate the quality of 360 videos in [7]. However, it is well-known that projected 2D images have redundancy due to over-sampling in certain areas. To deal with this problem, the concept of spherical PSNR (S-PSNR) is introduced in [9] to evaluate the quality of 360 video. In this metric, the points for PSNR calculation are obtained from unit spheres rather than 2D images.

In the latest stage of standardization for 360 videos [10], over ten PSNR-related objective quality measures should be reported for evaluating any 360 video coding technique. Obviously, this is very complicated for researchers to present and compare coding/adaptation techniques for 360 videos. These measures belong to five basic types of objective quality metrics, including PSNR, Weighted to spherically uniform PSNR (WS-PSNR) [11], spherical PSNR without interpolation (S-PSNR-NN) [12], spherical PSNR with interpolation (S-PSNR-I) [9], and PSNR in Crasters Parabolic Projection (CPP-PSNR) [13]. Also, the quality measures of these objective quality metrics can be classified into three phases of 360 video processing. Phase 1 is between input video and output video of the codec, phase 2 is between source video and output video of the codec, and phase 3 is between source video and reconstructed video.

To the best of our knowledge, no previous studies have evaluated these various PSNR-related quality measures for 360 videos. Moreover, it is well-known that PSNR-related measures do not represent well human perceived quality. However, the use and investigation of advanced quality metrics such as structural similarity-related metrics and content-based quality metrics (e.g. [3], [14], [15]) are still very limited. In this work, we investigate both objective quality and subjective quality of 360 videos. The goals are to identify appropriate objective quality metrics and to understand the perceived quality range provided by existing 360 videos. A subjective test is conducted to investigate the subjective quality of 360 videos encoded at different encoding parameters. Objective quality metrics are then evaluated based on their complexity and correlations with subjective quality. Regarding objective quality metrics, we not only consider the above PSNR-based metrics, but also investigate three other advanced metrics, namely, SSIM [14], MS-SSIM [15], and a hybrid metric using content features (HMCF) [3]. Based on the evaluation results, good objective quality metrics for different processing phases are identified. The tradeoffs of the quality metrics are also analyzed.

The remainder of the paper is organized as follows. In Sect. 2, an overview of 360 video and quality metrics is presented. Section 3 describes the details of the experiment. Section 4 discusses the obtained quality scores and relationship between objective quality measures and subjective quality. Finally, Sect. 5 concludes the paper and provides an outlook on future work.

Manuscript received April 5, 2017.
Manuscript revised August 26, 2017.
Manuscript publicized October 16, 2017.
[†]The authors are with the University of Aizu, Aizuwakamatsu-shi, 965–8580 Japan.
[††]The authors are with Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi, Vietnam.
a) E-mail: tranhuyen1191@gmail.com
DOI: 10.1587/transinf.2017MUP0011

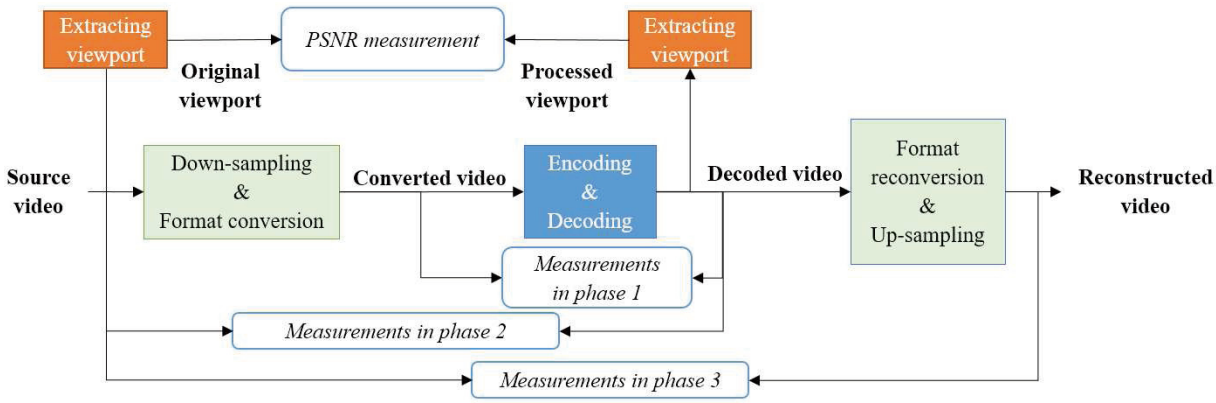


Fig. 1 Processing chain of 360-degree video.

2. Overview of 360 Video Processing and Quality Metrics

2.1 General 360 Video Processing

The general processing stages of a 360 video are illustrated in Fig. 1. A source 360 video in the source projection format is firstly down-sampled and/or converted to another projection format. After encoding-decoding, the decoded video is reconverted to the source projection format and/or up-sampled to the original resolution for quality evaluation. Although a 360 video is provided in every direction, a viewer sees only one direction at a time. Therefore, viewports extracted from videos could be used as an input for calculating objective quality.

To evaluate the impact of each stage on user's perception, various objective quality metrics could be considered. As recommended in [10], we evaluate quality measures in three phases. Phase 1 includes coding stage only, phase 2 includes coding stage and format conversion stage, and phase 3 includes all processing stages. Note that all metrics considered in [10] are PSNR and PSNR variants. As a viewport is extracted as a rectilinear image, its quality metric is PSNR only. We can see that quality measures in phase 1 quantify the impacts of encoding only. The measures in phase 2 quantify the impacts of distortions caused by down-sampling, forward projection format conversion, and encoding. Meanwhile, the measures in phase 3 also cover the impacts of reconversion to the source format.

In this paper, we consider eight quality metrics, which can be divided into three categories including PSNR-related metrics, structural similarity related metrics, and content-related metrics. The first category is based on the conventional PSNR, the second includes SSIM and MS-SSIM, and the third takes into account content characteristics.

2.2 PSNR-Related Metrics

In this work, there are five PSNR-related metrics, which are agnostic of the video content, as follows.

PSNR: this conventional measure is calculated based on the squared value differences of all points (or samples) between an original image and a test image with equal weights. So far, PSNR has been the de-facto quality metric in image and video coding.

WS-PSNR [11]: Weighted to Spherically uniform PSNR is calculated based on the squared value differences of all points between an original image and a test image, where the weight of each point depends on the sampling area on corresponding spherical surface. Similar to PSNR, WS-PSNR is also only used for two images of the same resolution and the same projection type.

S-PSNR-NN and S-PSNR-I: Spherical PSNR, which is first presented in [9], is calculated based on the squared value differences of points uniformly sampled on two conceptual unit spheres, one generated by an original image and one by a test image. In this way, two (panoramic) images with different resolutions and projection types can be compared. When the position on a unit sphere is rounded to the nearest neighbor position on the corresponding image, the metric is called S-PSNR-NN [12]. When the signal value of the position on the unit sphere is inferred by interpolation from neighbor positions on the corresponding image, the metric is denoted by S-PSNR-I.

CPP-PSNR [13]: To compute this PSNR-related metric, the original image and the test image are first converted into the Crasters Parabolic Projection format. This is similar to converting to the unit sphere, and so CPP-PSNR can be used on two images with different resolutions and projections.

2.3 Structural Similarity Related Metrics

Although PSNR is the simplest and most widely used quality metric, it is not very well matched to perceptual video quality [14]. An interesting type of quality metrics based on the concept of structural similarity is presented in [14], which can take advantage of characteristics of the human visual system and thus give a better correlation to perceptual video quality than PSNR. Two metrics of this type investigated in this work are as follows.

Table 1 Features of source videos.

Video	Motion activity	Spatial Complexity	Shooting type	Description
Video #1	Low	Complex	Static shooting	Natural video of game show genre. Movements of 4 characters in a room. Camera is fixed in floor.
Video #2	Medium	Simple	Dynamic shooting	Natural video of documentary genre. Dolphins move around in the ocean. Camera is controlled by a diver in a medium motion.
Video #3	Fast	Complex	Dynamic shooting	Natural video of adventure genre. Camera is in a roller coaster in a fast motion.
Video #4	Low	Complex	Static shooting	Natural video of documentary genre. Movements of panda bears in a National Nature Reserve.
Video #5	Fast	Complex	Dynamic shooting	Animated video of game genre. A gamer is playing Counter-strike Game.
Video #6	Medium	Medium	Static shooting	Animated video of animation genre. Movements of 1 cartoon character in a room.

Structural similarity (SSIM) [14] is calculated based on comparisons of luminance, contrast, and structure between two image signals x and y. Let μ_x and σ_x^2 be the mean and the standard deviation of x, respectively. Denote σ_{xy} the covariance of x and y. The comparison measures of luminance, contrast, and structure are respectively defined by

$$\mathrm{l}(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{1}$$

$$\mathrm{c}(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{2}$$

$$\mathrm{s}(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{3}$$

where C_1, C_2, and C_3 are model parameters. Finally, the SSIM metric between x and y is given by

$$SSIM(x,y) = [l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma}, \tag{4}$$

where α, β, and γ are parameters used to adjust the relative importance of the three components. The expression is simplified by setting $\alpha = \beta = \gamma = 1$ and $C_3 = \frac{C_2}{2}$ following [14]. Therefore, the SSIM metric is simplified as follows

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{5}$$

Multi-scale SSIM (MS-SSIM) [15] is calculated based on similar measures computed at different resolutions (or multi-scales). Specifically, the source image and the test image are firstly low-pass filtered, and then down-sampled by a factor of 2. The source image and the test image are denoted as scale 1. Denote M the highest scale which is obtained after $M-1$ interactions. For each scale j, the comparison measures of luminance $l_j(x,y)$, contrast $c_j(x,y)$, and structure $s_j(x,y)$ are calculated by Eqs. (1)~(3). Finally, the MS-SSIM metric is given by

$$\text{MS}-SSIM(x,y) = [l_M(x,y)]^{\alpha_M} \times \prod_{j=1}^{M}\left(\left[c_j(x,y)\right]^{\beta_j}\left[s_j(x,y)\right]^{\gamma_j}\right), \tag{6}$$

where α_M, β_j, and γ_j are parameters to define the relative importance of the different components. We can see that MS-SSIM includes a SSIM measure for scale M. In other words, if only the parameters α_M, β_M, and γ_M are non-zero values, MS-SSIM is equal to SSIM for scale M. In order to simplify the expression, the parameters are set as follows: $\alpha_j = \beta_j = \gamma_j$ with all scales $j \in \{1, 2, \ldots M\}$ and $\sum_{j=1}^{M}\gamma_j = 1$ following [15].

2.4 Content-Related Metric

In several previous studies, it is indicated that the impacts of content types (or content characteristics) on perceptual video quality are significant [2], [16]. However, there has been few proposed quality metrics considering the impacts of content characteristics. In this study, a content-related metric that is proposed in [3] is investigated.

Hybrid metric using Content Features (HMCF) [3] is calculated based on measured mean opinion score Q_0 (subjective quality metric) of an original video and PSNR (objective quality metric) between the original video and a test video. Specifically, HMCF is calculated by

$$HMCF = 1.04 \times Q_0\left(1 - \frac{1}{1 + e^{p(PSNR-s)}}\right), \tag{7}$$

where s and p are model parameters. Parameter s can be calculated by

$$s = \alpha - \beta * G_m - \gamma * NFD, \tag{8}$$

where α, β, and γ are model parameters, and NFD and G_m are parameters used to characterize content features. Specifically, NFD is used to measure the contrast of the video, and G_m is used to measure the amount of details in the video. More information about calculating NFD and G_m can be found in [3]. It should be noted that calculating NFD and G_m are very complex, and so the parameters for each content should be obtained in advance.

2.5 Selected Quality Measures

As mentioned, we evaluate quality measures in three phases (Fig. 1). The objective quality measures in phase 1 include PSNR, WS-PSNR, and HMCF. Note that the projection formats as well as the resolutions of the two videos must be identical. The quality measures in phase 2 include CPP-PSNR, S-PSNR-NN, and S-PSNR-I. Here the projection formats as well as the resolutions of the two videos can be different. For phase 3, the quality measures include the WS-PSNR, CPP-PSNR, S-PSNR-NN, SPSNR-I, PSNR, SSIM, MS-SSIM and HMCF. These measures are calculated between the source video and the reconstructed video with the same projection format and the same resolution.

In this study, we focus on comparing objective quality measures, and so only the most popular projection type, which is ERP, is used. In addition, because the video processing is omnidirectional, the video quality is almost identical across different viewing directions. Therefore, the PSNR measure of viewports is not considered in this study. Viewport PSNR evaluation, which is important when the quality is not omnidirectional, will be reserved for our future work.

So totally 14 objective quality measures (three in phase 1, three in phase 2, and eight in phase 3) are investigated in this paper. Note that all the PSNR-related measures described here are also required in [10].

3. Experiment Description

For the experiments, six 360 videos of 30-second duration with different levels of motion activity and spatial complexity are chosen. The characteristics of the videos are shown in Table 1. The test video streams are encoded by using H.264/AVC (libx264) with a frame rate of 30 fps. A GoP structure of "IBBP" with a GoP size of 30 is used for all videos. For each video, 20 encoding settings corresponding to combinations of five QP values of 22, 28, 32, 36, 40, and four resolutions of 3840x1920 pixels, 2880x1440 pixels, 2160x1080 pixels, 1440x720 pixels are used to generate 20 different video streams. Totally, there are 120 streams generated from the six original videos. Note that these resolutions and coding format are common in existing streaming platforms.

The objective quality measures are computed on an Ubuntu 14.04LTS PC with Intel Core i7 2.93GHz CPU and 8G RAM. PSNR and PSNR variants are deployed using 360Lib software package [17]. SSIM, MS-SSIM, and HMCF are implemented and added to 360Lib software package also. The quality measure of a video stream is the average of its frame quality values.

To evaluate image quality, it is shown that applying the SSIM measure locally is better than globally [14]. Therefore, in this study, SSIM measure is applied locally for sliding windows that move pixel-by-pixel across the whole image as in [14]. Then, SSIM index of the whole image is the mean of SSIM indexes of all windows in the image. In addition, to avoid "blocking artifacts" in the quality map, an 11x11 circular-symmetric Gaussian weighting function $\boldsymbol{w} = \{w_i | i \in \{1, 2, \ldots, N\}, \sum_{i=1}^{N} w_i = 1\}$ with the standard deviation of 1.5 samples is used as a smooth windowing approach [15]. μ_x, σ_x, and σ_{xy} are then modified as follows:

$$\mu_x = \sum_{i=1}^{N} w_i x_i \tag{9}$$

$$\sigma_x = \left(\sum_{i=1}^{N} w_i (x_i - \mu_x)^2 \right)^{1/2} \tag{10}$$

$$\sigma_{xy} = \sum_{i=1}^{N} w_i (x_i - \mu_x)(y_i - \mu_y). \tag{11}$$

For calculating SSIM, we set $C_1 = 6.5025$ and $C_2 = 58.5225$ following [14]. Regarding MS-SSIM metric, we use 5 scales and set $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$, and $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$ following [15].

In the subjective experiment, video streams are displayed by a device set consisting of a Samsung Galaxy S6 smartphone and a Samsung Gear VR headset. The point of view can be changed by moving viewer's head. The field of view of Samsung Gear VR is 96 degrees [18]. Samsung Galaxy S6 has the screen resolution of 2560x1440 pixels and the display size of 5.1 inches.

The subjective experiment is divided into two rounds. The first round is for the first three videos, and the second round for the last three videos (Table 1). In this study, mean opinion score (MOS) is used as the subjective quality metric for 360 videos. Specifically, each of the test streams is randomly displayed during the experiment, and then each viewer gives a rating score at the end of each stream with the score ranging from 1 (bad) to 5 (excellent). As 360 videos are new to many people, the viewers are required to watch some 360 videos one week in advance using available devices in our laboratory. In addition, in the experiment, before doing actual subjective tests, the viewers are trained to get accustomed to the devices and the rating procedure. In particular, the participants are trained by 5 training streams, which are different from the test streams. These training streams are displayed in the order from the best quality to the worst quality following the explanations and demonstrations of impairments caused by an increased QP and/or a decreased resolution. In the 5 training streams, one has the best quality, one has the worst quality, one demonstrates the impact of the QP increase, one shows the impact of the resolution decrease, and the last illustrates the impacts of both QP and resolution. During the experiment, every 20 minutes, there is a break for the viewers. There are totally 18 people taking part in the first round and 19 people in the second round of this experiment. The participants have ages between 20 and 37 with an average age of 25. The Absolute Category Rating method is used in our experiment [19]. A screening analysis of the subjective test results is performed

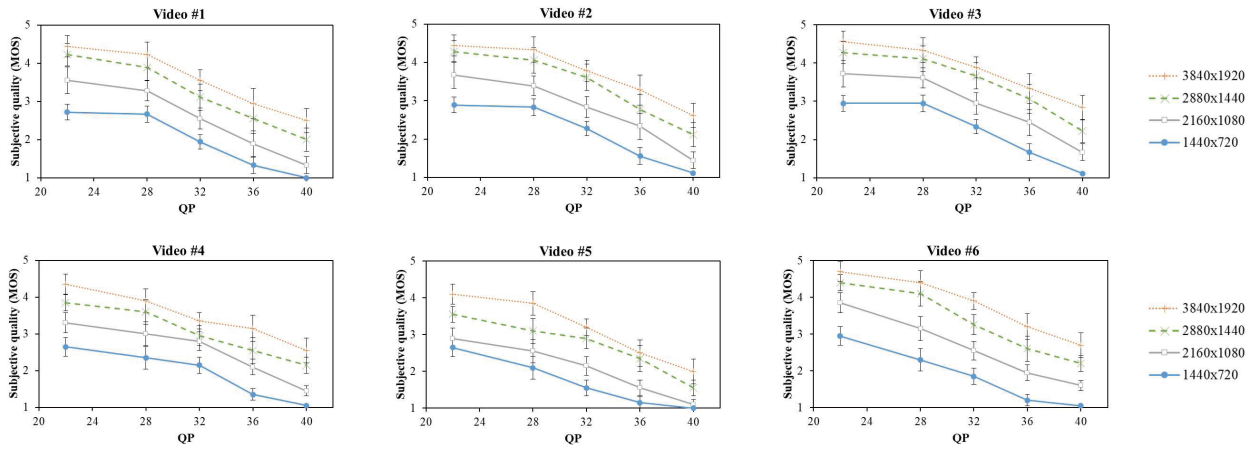
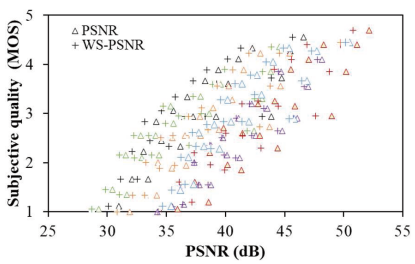
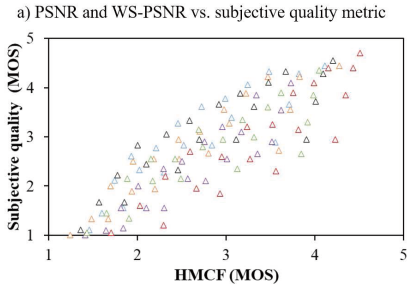


Fig. 2 Subjective quality values vs. encoding parameters.



a) PSNR and WS-PSNR vs. subjective quality metric



b) HMCF vs. subjective quality metric

Fig. 3 Coding distortion measurement. a) PSNR & WS-PSNR b) HCMF.

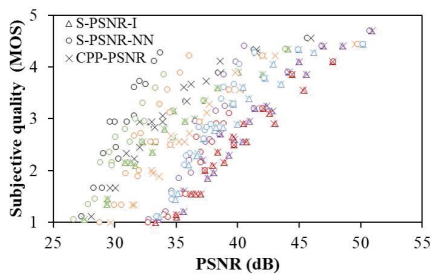
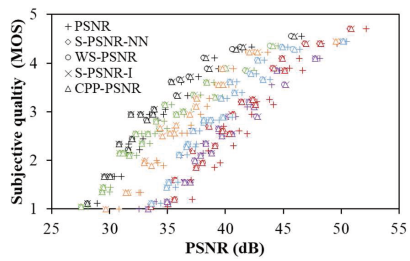
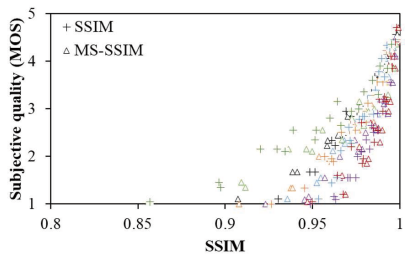


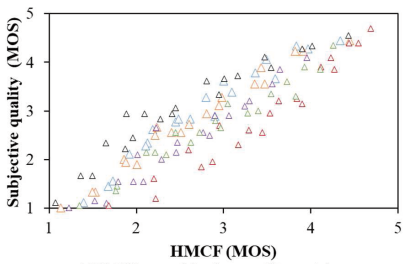
Fig. 4 Cross-format distortion measurement.



a) PSNR-related metrics vs. subjective quality metric



b) SSIM and MS-SSIM vs. subjective quality metric



c) HMCF vs. subjective quality metric

Fig. 5 End-to-end distortion measurement a) PSNR-related metrics b) SSIM and MS-SSIM c) HCMF

according to [19], and no subject is rejected.

4. Result Analysis

Figure 2 shows the subjective quality values (and corresponding confidence intervals) at different QP values and different resolutions for all considered videos. The relationship of each objective quality measure and the subjective quality is shown in Figs. 3, 4, and 5. In the following, we will discuss the characteristics of these measures in detail.

4.1 Subjective Quality

It can be seen from Fig. 2 that subjective quality values vary from 1 to about 4.5 (MOS). The maximum subjective quality values of Video #1, Video #2, Video #3, Video #4, Video #5, and Video #6 are 4.44, 4.44, 4.56, 4.35, 4.10, and 4.70 respectively. However, the quality drops quickly when QP

Table 2 Correlation coefficient between objective quality measures and subjective quality measure.

Objective quality measures		Video #1		Video #2		Video #3		Video #4		Video #5		Video #6		All videos	
		PCC	*RMSE*	*PCC*	*RMSE*	*PCC*	*RMSE*	*PCC*	*RMSE*	*PCC*	*RMSE*	*PCC*	*RMSE*	*PCC*	*RMSE*
Phase 1	WS-PSNR	0.93	0.37	0.91	0.40	0.89	0.42	0.89	0.40	0.92	0.34	0.90	0.46	0.75	0.64
	PSNR	0.93	0.37	0.91	0.40	0.90	0.42	0.89	0.40	0.92	0.35	0.90	0.47	0.76	0.63
	HMCF	0.92	0.38	0.89	0.43	0.88	0.45	0.87	0.42	0.92	0.36	0.90	0.47	0.85	0.51
Phase 2	CPP-PSNR	0.99	0.13	0.99	0.14	0.99	0.11	0.98	0.18	0.98	0.17	0.98	0.19	0.78	0.61
	S-PSNR-NN	0.92	0.39	0.98	0.18	0.94	0.33	0.95	0.27	0.96	0.26	0.94	0.35	0.67	0.72
	S-PSNR-I	0.99	0.13	0.99	0.14	0.99	0.11	0.98	0.18	0.98	0.17	0.98	0.19	0.78	0.60
Phase 3	WS-PSNR	0.99	0.14	0.99	0.13	0.99	0.12	0.98	0.17	0.98	0.16	0.98	0.19	0.78	0.62
	CPP-PSNR	0.99	0.14	0.99	0.13	0.99	0.12	0.98	0.17	0.98	0.17	0.99	0.18	0.78	0.61
	S-PSNR-NN	0.99	0.13	0.99	0.13	0.99	0.11	0.98	0.17	0.98	0.17	0.98	0.18	0.78	0.61
	S-PSNR-I	0.99	0.13	0.99	0.13	0.99	0.12	0.98	0.17	0.98	0.17	0.99	0.18	0.78	0.61
	PSNR	0.99	0.13	0.99	0.15	0.99	0.13	0.98	0.17	0.97	0.21	0.98	0.20	0.80	0.59
	SSIM	0.99	0.17	0.98	0.21	0.98	0.21	0.97	0.21	0.98	0.18	0.97	0.27	0.80	0.58
	MS-SSIM	0.98	0.18	0.98	0.17	0.99	0.15	0.97	0.22	0.97	0.21	0.97	0.26	0.87	0.48
	HMCF	0.99	0.14	0.99	0.16	0.99	0.15	0.98	0.18	0.97	0.21	0.98	0.20	0.91	0.41

is increased or the resolution is decreased.

Compared to the highest resolution of 3840x1920 pixels, the subjective quality of the resolution of 2880x1440 pixels is reduced by 0.4 MOS on average for all considered videos. When the resolution decreases further to 2160x1080 pixels and then 1440x720 pixels, the average quality reductions corresponding to these changes are respectively 0.61 MOS and 0.60 MOS. So, when the resolution is decreased from the highest level to the lowest level, the average quality degradation is about 1.61 MOS. Especially, although when the QP value is very good (QP = 22), subjective quality values at the resolution of 1440x720 pixels are lower than 3 MOS for all videos. This means that videos encoded at the resolution of 1440x720 pixels, which are being provided in current streaming platforms, are very negative to viewers. This is because, using the headset, viewers actually see roughly one sixth of the provided resolution on a large projected sphere. Therefore, 360 videos should be encoded at resolutions higher than 1440x720 pixels.

Similarly, at the highest resolution of 3840x1920 pixels, subjective quality values of video streams encoded at the QP of 40 are also lower than 3 MOS for all video types. This means that 360 videos should be encoded at QP values lower than 40. In addition, for most of the considered videos, to achieve subjective quality values higher than 3 MOS, the maximum QP values at the resolutions of 3840x1920 pixels, 2880x1440 pixels, and 2160x1080 pixels are 32, 32 and 28, respectively. It should be noted that these QP values are specific to the AVC coding format.

4.2 Quality Correlation for Individual Videos

In this part, we will investigate the correlation of different objective quality measures with subjective quality for each video. In Figs. 3, 4, and 5, each marker shows a type of objective quality metric and each color corresponds to a video (i.e., orange for Video #1, blue for Video #2, black for Video #3, green for Video #4, purple for Video #5, and red for Video #6).

To investigate the relationships between objective quality and subjective quality, different mapping functions could be used, e.g. linear, exponential, power, and logistic functions. After trying curve-fitting with these functions, it is found that the Pearson Correlation coefficients (PCCs) of the logistic function are always highest. Specifically, the average PCCs (over all metrics and videos) of linear, exponential, power, and logistic functions, are 0.93, 0.90, 0.92, and 0.97, respectively. Therefore, in this study, a four-parameter logistic function of the form

$$f(x) = d + \frac{a-d}{1+\left(\frac{x}{c}\right)^b} \tag{12}$$

is used to map between the objective quality values and the subjective quality values. Note that all model parameters including a, b, c, d, α, β, γ and p in Eqs. (7), (8), and (12) are obtained by curve-fitting using the data in Figs. 3, 4, and 5.

The correlation coefficients including Pearson Correlation Coefficient (PCC) and Root Mean Square Error (RMSE), which are used to quantify how well the objective quality and subjective quality correlate, are shown in Table 2. Note that the last two columns of Table 2 will be reserved for the next subsection, which discusses the correlation for all videos.

From Table 2, we can see that although the six videos used in this study have different characteristics such as motion activity, spatial complexity, genre, and shooting type, the behaviors of PCC and RMSE are consistent for all six videos. Specifically, the objective quality measures used in phase 1 have rather low PCC (0.87~0.93) and rather high RMSE (0.34~0.47).

In phase 2, the objective quality measures CPP-PSNR and S-PSNR-I have very high PCC (0.98~0.99) and very low RMSE (0.11~0.19). Meanwhile, the S-PSNR-NN measure in phase 2 has lower PCC (0.92~0.98) and higher RMSE (0.18~0.39). This means that in phase 2, CPP-PSNR and S-PSNR-I are closer to users' perception than S-PSNR-NN. Especially, the correlation coefficients of S-PSNR-NN are different for different videos. This implies that the accuracy of rounding using the nearest neighbor position in S-PSNR-NN depends on content features. For example, be-

cause the background of Video #2 is very simple, the rounding using the nearest neighbor position in this video causes less errors than in the other videos.

It should be noted that in phase 2, the resolution of the decoded video is equal or lower than that of the source video. The smaller the resolution of the decoded video is, the more significant the impact of rounding is. However, in phase 3, as the resolution of the reconstructed video is already upsampled to be equal to that of the source video, the PCC of S-PSNR-NN measure for each video is always high (0.98~0.99).

In phase 3, PSNR, PSNR variants, and HMCF have very high PCC (0.97~0.99) and very low RMSE (0.11~0.21). Interestingly, it turns out that the traditional PSNR, the most straightforward calculation, has high PCC similarly to those of the other measures in phase 3. So PSNR in phase 3 could be directly used for 360 videos.

Meanwhile, SSIM and MS-SSIM have a little lower PCC (0.97~0.99) and higher RMSE (0.15~0.27). Especially, for Video #1 and Video #5, PCC of MS-SSIM is lower than that of SSIM. Meanwhile, for Video #2 and Video #3, the correlation coefficient of MS-SSIM is better than that of SSIM. It be because that the SSIM and MS-SSIM metrics have parameters used to adjust the relative importance of different components. The parameters are related to the human visual system, and so it is difficult to directly obtain them from simple subjective experiments [15]. In addition, we can see that the RMSE of the HMCF measure in phase 3 is a bit higher (or worse) than that of PSNR-related measures for each video.

Therefore, the objective measures including CPP-PSNR and S-PSNR-I in phase 2 and PSNR-related measures in phase 3 are not only effective but also less complicated to evaluate the quality.

4.3 Quality Correlation for All Videos

Though the quality measures in phase 2 and phase 3 have high correlations for each individual video, they mostly have low PCC and high RMSE when fitting for all six videos (see the last two columns of Table 2). In fact, only HMCF measure in phase 3 has acceptable results.

More specifically, for all videos, only the HMCF measure in phase 3 still has high PCC (i.e., 0.91). It is because that HMCF is calculated based on some of content features including NFD and G_m as presented in Sect. 2. Therefore, HMCF metric in phase 3 is able to compare 360 video quality across different videos. However, we can see that the RMSE of HMCF measure for all videos is up to 0.41 MOS. This means that it is necessary to improve content-related metrics.

As for SSIM and MS-SSIM measures, it is interesting that the PCC of SSIM is not good, while the PCC of MS-SSIM is worse than that of HMCF only. This suggests that MS-SSIM is the second choice to compare quality values across different videos.

Table 3 Time complexity of objective quality metrics for two frames with the same resolution of 3840x1920 pixels.

Objective quality metrics	Time complexity (ms)
PSNR	19
S-PSNR-NN	223
WS-PSNR	44
S-PSNR-I	879
CPP-PSNR	7979
SSIM	394
MS-SSIM	581
HMCF	19

4.4 Complexity Evaluation

Table 3 shows the complexity measured as the amount of time taken to calculate each objective quality metric. For fair comparison (i.e. with same resolution and same projection), the metrics are compared in phase 3 only. Note that, as mentioned, SSIM, MS-SSIM, and HMCF metrics require complex (offline) processes to obtain content-related parameters in advance. However, these processes are not reflected in these complexity values.

We can see that PSNR and HMCF have the smallest time complexity. Compared to PSNR, WS-PSNR takes about two times longer to calculate. That is due to the additional time to process the weight of each sample. In addition, the complexity values of S-PSNR-NN and S-PSNR-I are about eleven times and forty times higher than that of PSNR. It is because rounding by using the nearest neighbor position takes a shorter time than taking interpolation. Also, the complexity of CPP-PSNR is about four hundred times higher than that of PSNR. It is because converting to Craster parabolic projection is rather complex. The complexity of SSIM is twenty times higher than that of PSNR. Meanwhile, MS-SSIM, with multiple rounds of down-sampling, takes about thirty times longer to calculate compared to PSNR.

Therefore, in phase 3, given the smallest complexity and the high correlation with subjective quality, PSNR is the most appropriate measure for evaluating different streams of an individual video. However, for evaluating video quality across different videos, HMCF is the most appropriate metric for 360 video communications.

4.5 Remarks on Evaluation Results

Based on the above results and discussions, some remarks on the quality metrics can be summarized as follows.

- To achieve acceptable subjective video quality, 360 videos should be encoded at resolutions higher than 1440x720 pixels, regardless of QP values.
- The quality measures in phase 1 cannot be used to predict user perceived quality of 360 videos.
- In phase 2, quality measures CPP-PSNR and S-PSNR-I, but not S-PSNR-NN, are appropriate to evaluate 360 video quality.
- In phase 3, actually all quality measures are well correlated with subjective quality. Especially, the tradi-

tional PSNR, which has the simplest calculation, provides very high correlation results. Among these measures, SSIM and MS-SSIM have a little lower performance and more complex processing. So, SSIM and MS-SSIM should not be used as a quality metrics for 360 video communications.

- Most quality measures are only good to compare video quality between streams of the same video. To evaluate the quality across different 360 videos, HMCF in phase 3 should be used.
- Content-related quality metrics have content-related parameters, which are not easy to obtain in advance. This is the main drawback of the content-related metrics. Such content-related parameters can be obtained by computing content features (e.g. spatial information) [20].

5. Conclusions

In this paper, we have investigated the objective and subjective quality for 360 videos. The subjective results showed that the quality range of existing 360 videos spans from 1 MOS to 4.5 MOS. In addition, to achieve acceptable video quality, 360 videos should be encoded at resolutions higher than 1440x720 pixels. An investigation of the relationships between objective quality metrics and subjective quality metric was also conducted. Totally fourteen objective quality measures were considered in this paper. Various tradeoffs of these metrics were identified in the evaluation. It was found that most of the PSNR-related quality measures are well correlated with subjective quality. However, for evaluating video quality across different contents, a content-based quality metric is needed. For future work, quality evaluation for adaptive streaming of 360 videos will be investigated.

References

[1] H.T.T. Tran, N.P. Ngoc, A.T. Pham, and T.C. Thang, "A multi-factor QoE model for adaptive streaming over mobile networks," 2016 IEEE Globecom Workshops (GC Wkshps), Washington, DC, pp.1–6, Dec. 2016.

[2] H.T.T. Tran, N.P. Ngoc, Y.J. Jung, A.T. Pham, and T.C. Thang, "A Histogram-Based Quality Model for HTTP Adaptive Streaming," IEICE Trans. Fundamentals, vol.E100-A, no.2, pp.555–564, Feb. 2017.

[3] Y.-F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," IEEE Trans. Circuits Syst. Video Technol., vol.21, no.3, pp.286–298, Oct. 2010.

[4] A. Steed, S. Frlston, M.M. Lopez, J. Drummond, Y. Pan, and D. Swapp, "An 'In the Wild' Experiment on Presence and Embodiment using Consumer Virtual Reality Equipment," IEEE Trans. Vis. Comput. Graphics, vol.22, no.4, pp.1406–1414, Jan. 2016.

[5] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, "An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments," 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, pp.1–6, June 2016.

[6] H.T.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 International Workshop on Multimedia Signal Processing, London-Luton, UK, Oct. 2017.

[7] F. Qian, L. Ji, B. Han, and V. Gopalakrishnan, "Optimizing 360 video delivery over cellular networks," Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges, New York, USA, pp.1–6, Oct. 2016.

[8] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," 2017 IEEE International Conference on Communications (ICC), Paris, France, pp.1–7, May 2017.

[9] M. Yu, H. Lakshman, and B. Girod, "A Framework to Evaluate Omnidirectional Video Coding Schemes," 2015 IEEE International Symposium on Mixed and Augmented Reality, Fukuoka, pp.31–36, Sept. 2015.

[10] Y. Ye, E. Alshima, and J. Boyce, "JVET-E1003: Algorithm descriptions of projection format conversion and video quality metrics in 360Lib," Joint Video Exploration Team (JVET) of ITU-T SG 16 WP3 and ISO/IEC JTC 1/SC 29/WG 11 5th Meeting, Geneva, 2017.

[11] Y. Sun, A. Lu, and L. Yu, "AHG8: WS-PSNR for 360 video objective quality evaluation," Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0040, 4th Meeting, Chengdu, 2016.

[12] Y. He, B. Vishwanath, X. Xiu, and Y. Ye, "AHG8: InterDigital's projection format conversion tool," Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0021, 4th Meeting, Chengdu, 2016.

[13] V. Zakharchenko, E. Alshina, A. Singh, and A. Dsouza, "AHG8: Suggested testing procedure for 360-degree video," Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0027, Chengdu, 2016.

[14] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process., vol.13, no.4, pp.600–612, April 2004.

[15] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, California, pp.1398–1402, Nov. 2003.

[16] T.C. Thang, J.W. Kang, and Y.M. Ro, "Graph-based Perceptual Quality Model for Audiovisual Contents," 2007 IEEE International Conference on Multimedia and Expo, Beijing, pp.312–315, July 2007.

[17] ITU-T/ISO/IEC Joint Video Exploration Team, "360Lib," https://jvet.hhi.fraunhofer.de/svn/svn_360Lib/tags, accessed July 20, 2017.

[18] VR Times, "Comparison Chart of FOV (Field of View) of VR Headsets," http://www.virtualrealitytimes.com/2015/05/24/chart-fov-field-of-view-vr-headsets/, accessed July 20, 2017.

[19] R. ITU-T P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment," 2014.

[20] P. Paudyal, F. Battisti, and M. Carli, "Impact of video content and transmission impairments on quality of experience," Multimedia Tools and Applications, vol.75, no.23, pp.16461–16485, Dec. 2016.

Huyen T. T. Tran received the B.E. degree from Hanoi University of Science and Technology in 2014. She is currently a graduate student and research assistant in the Computer Communications Lab., the University of Aizu. Her research interests include Quality of Experience (QoE), multimedia networking, and content adaptation. She is a recipient of Japanese government scholarship (MonbuKagaku-sho) for graduate study since 2015.

Cuong T. Pham is currently a student and research assistant in Embedded System and Reconfigurable Computing Lab., Hanoi University of Science and Technology, Vietnam. His research interests include multimedia networking and content adaptation.

Nam Pham Ngoc received B.E. degree in Electronics and Telecom. from Hanoi University of Science and Technology (Vietnam) and M.Sc. degree in Artificial Intelligence from K.U. Leuven (Belgium) in 1997 and 1999, respectively. He was awarded a Ph.D. degree in Electrical Engineering from K.U. Leuven in 2004. From 2004 until now he has been working at Hanoi University of Science and Technology, Vietnam. His research interests include QoS management at end-systems for multimedia applications, reconfigurable embedded systems and low-power embedded system design.

Anh T. Pham received the B.E. and M.E. degrees, both in Electronics Engineering from the Hanoi University of Technology, Vietnam in 1997 and 2000, respectively, and the Ph.D. degree in Information and Mathematical Sciences from Saitama University, Japan in 2005. From 1998 to 2002, he was with the NTT Corp. in Vietnam. Since April 2005, he has been on the faculty at the University of Aizu, where he is currently Professor and Head of Computer Communications Laboratory with the Division of Computer Engineering. Professor Pham's research interests are in the broad areas of communication theory and networking with a particular emphasis on modeling, design and performance evaluation of wired/wireless communication systems and networks. He has authored/co-authored more than 140 peer-reviewed papers, including 40+ journal articles, on these topics. Professor Pham is senior member of IEEE. He is also member of IEICE and OSA.

Truong Cong Thang received the B.E. degree from Hanoi University of Science and Technology, Vietnam, in 1997 and the Ph.D. degree from KAIST, Korea, in 2006. From 1997 to 2000, he worked as a network engineer in Vietnam Post & Telecom (VNPT). From 2007 to 2011, he was a Member of Research Staff at Electronics and Telecommunications Research Institute (ETRI), Korea. He also has been an active member of Korean and Japanese delegations to standard meetings of ISO/IEC and ITU-T since 2002. Since 2011, he has been an Associate Professor of University of Aizu, Japan. His research interests include multimedia networking, image/video processing, content adaptation, IPTV, and ISO/IEC/ITU standards.