Düzce University
Journal of Science & Technology

Research Article

Comparison of Subjective QoE Models for Auto-Stereoscopic 3DTV

Nükhet ÖZBEK[a,*]

[a] *Department of Electrical and Electronics Engineering, Faculty of Engineering, Ege University, İzmir, TURKEY*
** Corresponding author's e-mail address: nukhet.ozbek@ege.edu.tr*
DOI : 10.29130/dubited.489137

ABSTRACT

Measuring Quality of Experience (QoE) of stereoscopic 3D video is a hot research topic. Subjective models are considered as the most reliable and facilitate development of objective models. However, to collect user opinion scores takes long time. Therefore, new subjective assessment models should be proposed providing not only time-efficiency but also good accuracy and reliability. In this study, two novel subjective QoE models are proposed as alternative to the conventional Double Stimulus Continuous Quality Scale method. Also, a fair comparison method is proposed to evaluate performances of the three subjective methods using the same stimuli prepared with the most recent multi-view video codec on an auto-stereoscopic 3DTV. Correlations are calculated using two objective QoE measures using depth maps and structural similarities. The results demonstrate that the performances of the proposed models are comparable to each other and both models are superior to the conventional method.

Keywords: *QoE, 3DTV, SSIM, multi-view video plus depth, paired comparison*

Auto-stereoskopik 3BTV için Öznel Deneyim Kalitesi Modellerinin Karşılaştırılması

ÖZET

Stereoskopik 3B videonun deneyim kalitesini (DK) ölçmek güncel bir araştırma konusudur. Öznel modeller en güvenilir olarak düşünülür ve nesnel modellerin geliştirilmesine fayda sağlar. Ne var ki, kullanıcı görüş notlarını toplamak uzun zaman almaktadır. Bu nedenle, sadece zaman verimliliği değil aynı zamanda iyi netlik ve verimlilik sağlayan yeni öznel değerlendirme modelleri önerilmelidir. Bu çalışmada, geleneksel Çift Uyarıcılı Sürekli Kalite Ölçeği metoduna alternatif olarak iki yeni öznel DK modeli önerilmektedir. Bununla birlikte, bu üç öznel metodun performansını oto-stereoskopik 3BTV üzerinde değerlendirmek için en yeni çok-bakışlı video kodlayıcısı ile hazırlanmış aynı uyartanları kullanarak adil bir karşılaştırma yöntemi önerilmektedir. Korelasyonlar derinlik ve yapısal benzerlik haritalarını kullanan iki nesnel DK ölçüsü kullanılarak hesaplanmıştır. Sonuçlar önerilen modellerin performanslarının birbiri ile kıyaslanabilir ve her iki modelin de geleneksel metottan üstün olduğunu göstermiştir.

Anahtar Kelimeler: *Deneyim kalitesi, 3BTV, SSIM, çok-bakışlı video artı derinlik, ikili karşılaştırma*

Received: 28/11/2018, Revised: 25/06/2019, Accepted: 01/07/2019

I. INTRODUCTION

Rapid advances in 3D video compression and transmission technologies required new models for subjective and objective measurement of Quality of Experience (QoE). Video Quality Experts Group (VQEG) launched the 3DTV Project to research perceptual video quality assessment of 3DTV. To quantify the quality experienced by the users the subjective QoE methods are proposed. The subjective QoE measures are commonly employed as ground truth for evaluation of the objective QoE methods. Absolute Category Rating (ACR) and Double Stimulus Continuous Quality Scale (DSCQS) methods are traditional 2D subjective QoE models, described in ITU-R Recommendation BT.500 [1] to quantify mono-dimensional factors. However, it is not clear whether they are applicable to multi-dimensional subjective impression which composed of image and depth quality and visual comfort.

DSCQS is the widely used one, but it certainly has some issues. First of all, the assessment is a time-consuming process. Secondly, random order of presentation may cause the contextual effect. Last, sequential presentation of the videos results in weaker concentration and so false judgement of the assessor.

As an alternative to DSCQS, Subjective Assessment Methodology for Video Quality (SAMVIQ) is recommended in BT.1788 [2]. SAMVIQ is an interactive subjective model, where test videos are presented in Multiple Stimuli (MS) way to make the user can decide the order of test, and revise the ratings. SAMVIQ brings better reliability due to smaller deviation of user ratings. Özbek et al. adopted SAMVIQ in [3] for QoE measurement of asymmetrically coded stereo videos. Aflaki et al. searched in [4] the constraints which limit the preference of utilizing asymmetric coding compared to symmetric coding using systematic subjective testing.

Li et al. [5] chose the Paired Comparison (PC) methodology [6] to analyze the impact of two factors for 3DTV, such as the test environment and the display technology. PC was adopted in [7], as well. They stated that the assessor might have difficulty to assign an absolute psychophysical score for the stimulus since the viewer is not used to watch 3DTV, and has not enough experience. Providing preference between two stimuli is much easier than rating such stimuli for assessors.

Simultaneous Presentation (SP) is an alternative to continuous presentation which described in [1]. One advantage of the method is that SP brings considerable amount of reduction in the test duration due to simultaneous presentation of reference and test stimuli. Another is it enables easy evaluation of the differences between the stimuli.

Thus, new subjective assessment models should be proposed providing not only time-efficiency but also good accuracy and reliability [8, 9]. In [10], we proposed to extend DQCQS with the Multiple Stimuli plus Simultaneous Presentation, and called as MSSP-DSCQS. We demonstrated increased the results' reliability, enhanced time-efficiency and need of less assessors. In this work, we propose to employ ITU-T Recommendation P.910 [6], the PC methodology, which is capable of evaluating the distortions where the artifacts are difficult to discriminate. Also, we propose a fair comparison method to evaluate performances of different subjective QoE models using the most recent multi-view video codec on an auto-stereoscopic 3DTV. The rest of the paper is organized as follows: Section II is about auto-stereoscopic 3DTV. Section III gives information about the state-of-the-art multi-view video codec. Section IV presents the subjective QoE models proposed. Section V presents the objective QoE methods used. Section VI presents and discusses the experimental results. Section VII is the Conclusion.

II. AUTO-STEREOSCOPIC 3DTV

Auto-stereoscopic 3D televisions typically need capturing and transmission of 5 or 9 views. Due to huge amount of data, Multi-view Video plus Depth (MVD) format is developed which only includes 2 or 3 views along with corresponding per-sample depth data [11]. Thanks to depth data intermediate views between original camera positions are synthesized using Depth Image Based Rendering (DIBR) algorithms [12].

To show MVD content, we use Tridelity ML4210va in "3DTV Laboratory" of Ege University. Tridelity ML4210va is a 5-view 1920x1080p 42" wide auto-stereoscopic 3DTV that includes parallax barrier technology. To from five viewing zones the 3DTV requires five interleaved YUV files as input. Video sequences are encoded and reconstructed, and then virtual views are rendered with the MVD codec which is 3D extension of High Efficiency Video Coding (HEVC) [13]. Finally, five views are multiplexed using our own interleaving software.

The assessor is placed approximately 3.5 meters away from the 3DTV keeping the distance of 8H recommended for the laboratory conditions in ITU-R BT.2021 [14]. We selected Balloons, Kendo, Poznan Street and Undo Dancer test videos to prevent discomfort and fatigue in eyes because of 3D quality of raw (uncompressed) videos. We watched the original YUV files in the 3DLAB to understand if the videos perceived properly in terms of depth, color brightness, and visual angle. Fig. 1 gives the certain images of the MVD test videos that are selected to represent the content. The content of test sequences varies in amount of detail in spatial, motion, depth, and light features. They are explained as follows:

Balloons sequence has panning camera and medium object motion, complex texture and depth. Kendo sequence has also panning camera but high object motion, medium texture and depth. Undo Dancer sequence has medium object and complex camera motion, simple texture and depth. Poznan Street sequence has stationary camera, high motion objects, complex texture and depth.

Figure 1. *The certain images of the MVD test videos that are selected to represent the content*

III. MULTI-VIEW VIDEO COMPRESSION

3D extension of H.265/HEVC [15] is employed, and the reference software of 3D HEVC Test Model [16], i.e. the 3D-HTM codec, is used for MVD video encoding. Version 16.0 of the HTM reference software [17] is used in to compress the MVD stimuli in Table 1. Poznan Street is compressed using the two-view configuration. Other three sequences are compressed using the three-view configuration. Codec is set according to the common test conditions specified in [18]. We set the Group of Picture size and the Intra Period to 8 and 24, respectively. The 3-view encoding structure of the 3D-HTM is as follows: The middle view, i.e. View-3, is encoded independently as Intra-view whereas View-1 and View-5 are encoded as Predictive-view with inter-view prediction from View-3. View-2 and View-4 are virtual views synthesized using DIBR.

We employed QP parameter to change the fidelity such as 25, 30, 35, 40, and 45. To select three test points (QP values) we used the comparator test program which we developed in our previous study [3]. We visually test all of the QP options compressed to investigate the perceptually noticeable difference levels. The lowest and highest QoE levels are picked up as the test points, and labeled as “QP35” and “QP45”, respectively. After that, one more test point is picked up for the QoE level of that has slight difference compared to the previously selected test points, and named as “QP40”. Note that, there is no noticeable difference between subjective qualities of QP values of 25, 30, and 35. Thus, we have 4x3=12 MVD test sequences to be evaluated in total.

Table 1. *Properties of MVD stimuli*

Video	Resolution	Selected V+D	Owner
Balloons	1024x768	1-3-5	Nagoya U.
Kendo	1024x768	1-3-5	Nagoya U.
Undo Dancer	1920x1088	1-2-3	Nokia C.
Poznan Street	1920x1088	3-4	Poznan U.

IV. SUBJECTIVE QOE MODELS

A. THE DSCQS METHODOLOGY

In the conventional DSCQS methodology, pairs of videos, namely stimuli A and B, are sequentially presented to the assessor. The stimulus A is always the reference/original but the observer does not know it. The stimuli pairs are presented randomly for each assessor. The order of stimulus A/B in the pair is random, as well. For each video pair, the subject must watch the red and the green stimulus sequentially, without knowing which one is the tested and which one is the reference, twice. Then the observer vote the red and the green stimuli on the two sliders where values in 0-100 are grouped in 5 groups as bad, poor, fair, good, and excellent. They are asked to give a score the stimuli for QoE in sharpness, depth, and naturalness.

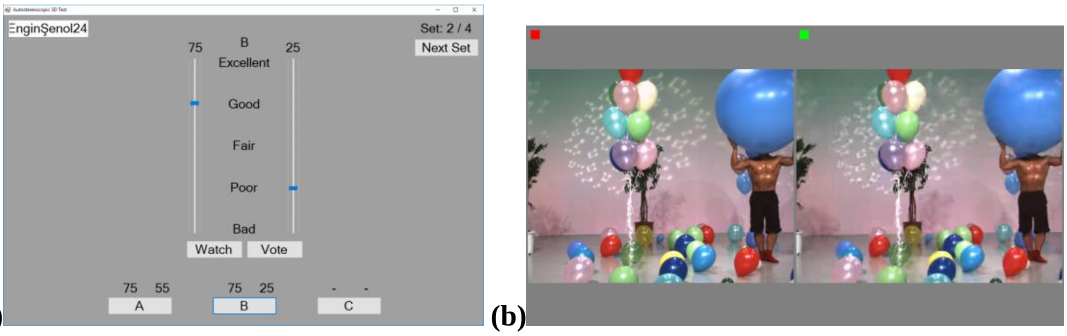
Difference Mean Opinion Score (DMOS) is employed as a QoE measure in DSCQS. First, the score given by the assessor to the tested stimulus is subtracted from the score given by the same assessor to the related reference stimulus and it is called as differential score. The differential scores are computed for each assessor and each test video. Then, they are normalized to values in 0-100. Finally, the DMOS value is calculated as the average of the normalized values of scores.

However, the DSCQS method has a few shortcomings: One issue is low consistency of scores which needs higher number of subjects to reach sufficient reliability. Inconsistency of scores may be explained as that higher QoE test may have a lower score than lower QoE test, since presentation order is random. Some assessors may be rejected due to inconsistent scores. Therefore, the need for more assessors and time is inevitable.

Another issue is the fact of “contextual effect” that stems from the random presentation order of stimulus in the pair. In case of lower QoE video of the pair shows up before the higher QoE video, actually the reference, the assessor may give lower scores to each stimulus of the pair and so he is also rejected because of inconsistent score for the reference.

B. THE MSSP-DSCQS METHODOLOGY

We implemented MSSP-DSCQS in [10] to get rid of the issues of DSCQS. MS form is inspired by the SAMVIQ methodology to increase reliability. SP is a solution for random and sequential presentation of stimuli pair in DSCQS. Fig. 2 gives screenshots of the MSSP-DSCQS methodology.



(a) **(b)**

Figure 2. (a) *Graphical user interface for MSSP-DSCQS.* ***(b)*** *Screenshot for simultaneous presentation of stimuli (red : the reference, green : the tested)*

The evaluation of one sequence starts whence the evaluation of another sequence is completed. In order words, all qualities of the current sequence must be scored to pass the next sequence. The quality access is randomized from one sequence to another to let the subjects vote without any effect caused by the button names. As mentioned earlier 4 different MVD sequences and 3 QoE levels are selected in setup phase. Fig. 2 (a) shows the evaluation process of the second video sequence (Set: 2 / 4) with the three qualities that are randomly assigned to the buttons A, B, and C.

The evaluation is carried out interactively which means the subject can watch and vote any test point in any order, together with revote and replay abilities. Thus issues of DSCQS can be resolved since the assessors have chance for direct comparison of the distorted videos with one another or against the reference (as shown in Fig. 2 (b)). The green stimulus is always the tested while the red stimulus is the reference, and the observers were told about it. Within the same sequence, all the subjects rated the red stimulus with the same score. Score of the red stimulus was generally equal or higher than score of the green stimulus so inconsistent scoring issue of DSCQS can be resolved. Similar to the SAMVIQ methodology, we used MOS (Mean Opinion Score) scoring instead of DMOS scoring. MOS is calculated by averaging the quality ratings collected from the assessors.

C. THE SP-PC METHODOLOGY

The PC methodology has been described in ITU-T Recommendation P.910 [6]. For example, the test stimuli of {A, B, C} are prepared as all the possible n(n-1)/2 combinations such as AB, AC, and BC. After each pair is presented, a judgment is made on which stimulus in a pair is preferred in the context of the test scenario.

The PC method is capable of assessing the distortions where the artifacts are hard to discriminate, and so are similar QoE levels. Advantages of the PC method are as follows: First, simplicity: In stimulus evaluation, a subject is asked for only the preference instead of a discrete or continuous quality rating. Second, reliability: Quantitative judgment leads to large variances across non-expert subjects and thus unreliable results. PC reduces complexity, and improves reliability.

To implement the PC methodology, we modified MSSP-DSCQS and named it Simultaneous Presentation of Pair Comparison (SP-PC). Since the subject can compare the impaired videos against each other simultaneously, the simplicity of paired comparison is further improved. Fig. 3 (a) gives user interface of the proposed test procedure, which is briefly explained as follows:

The evaluation process is performed where one pair is followed by another pair. The total number of pairs depends on the design of stimuli. In this study, the test includes comparison between different stimuli obtained using 3 different QP values for the same video, which makes 4x $\binom{3}{2}$=12 pairs. In Fig. 3 (b), two different qualities of Kendo sequence are shown during the comparison between QP40 and QP45.

While assessing the current pair, the subject can interactively play and judge, with ability to replay. Assessors can express their judgments by simply clicking one of the three buttons, where three discrete grades are represented as ‘Red is better’, ‘Green is better’, and ‘Same’. (See Fig. 3 (a)) From one test to another, presentation order of pairs is randomized to prevent the subjects score being biased.

(a) (b)

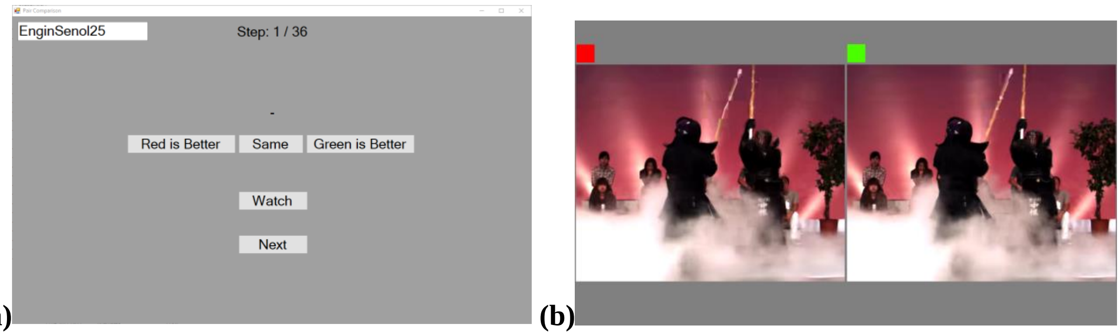


Figure 3. (a) *Graphical user interface for SP-PC.* ***(b)*** *Screenshot for simultaneous presentation of stimuli (red : QP45, green : QP40)*

On the other hand, PC brings some challenges in the test set-up and analysis of the test results such as: First, the number of pairwise comparisons increases exponentially with the number of stimuli under evaluation. Second, assessor preferences need to be translated into quality scores.

In [19, 20], Lee et al. propose solutions and systematic ways for these challenges. Based on comparison results, winning frequencies are calculated as the number of wins of a stimulus against the other stimuli for multiple assessors. Through data processing, they are translated into relative quality scores in each group.

Subjective tests result with matrixes that show comparative score of each stimulus in reference to other stimuli. Obtained matrixes consist of values, 1 or 0, only. By summing matrixes over all the assessors, overall comparative scores are achieved and used to calculate the quality score of each stimulus. Win and tie count matrixes for each video are obtained after subjective tests, and converted to relative quality scores for each stimulus using Bradley-Terry-Luce (BTL) model [21]. Relative quality scores are used to give a final statement about the subjective QoE of each stimulus. A BTL implementation in Matlab, OptiPt, developed in [22], is used in our work to obtain quality scores of stimuli.

V. OBJECTIVE QOE METHODS

We measure objective QoE of MVD content by using local and global quality weighting approaches developed earlier in [10]. It is also shown that our methods outperform the state-of-art method [23] which does not employ the depth information. In both methods, Structural SIMilarity (SSIM) [24] maps of the middle view (View-3), and its nearby view (View-4), are combined with depth maps since the assessors sit watching the middle zone of the 5 viewing zones during the visual tests. In [10], superiority of the proposed metrics is presented for constant QP MVD encodings. In this work, these metrics are used as objective QoE methods to compare correlations.

A. LOCAL DEPTH WEIGHTING

In the first method, depth distortion is locally computed as the Euclidian distance between every pixel of the uncompressed depth map and of the distorted depth map. M_{mapV3} and M_{mapV4} are the SSIM maps between the uncompressed and the distorted texture frames of View-3 and View-4, respectively. Ddl_{V3} and Ddl_{V4} are the combined measures, calculated by

$$Ddl_{V3}\,(p) = M_{mapV3}(p)\left(1-\frac{\sqrt{UncompressedDepth{-}V3(p)^2-DistortedDepth{-}V3(p)^2}}{255}\right) \quad (1)$$

$$Ddl_{V4}\,(p) = M_{mapV4}(p)\left(1-\frac{\sqrt{UncompressedDepth{-}V4(p)^2-DistortedDepth{-}V4(p)^2}}{255}\right) \quad (2)$$

Local Depth Combination of SSIM (LDC-SSIM) is computed by averaging over N pixels and over the views 3 and 4:

$$LDC\text{-}SSIM = \frac{1}{2}\left(\frac{1}{N}\sum_N Ddl_{V3}(p) \;+\; \frac{1}{N}\sum_N Ddl_{V4}(p)\right) \quad (3)$$

The final objective QoE measure is obtained by averaging Eq. 3 over all frames of the sequence.

B. GLOBAL DEPTH WEIGHTING

In the second method, Ddg, the global depth distortion is used as depth map combination. Ddg is obtained as the SSIM index between uncompressed and distorted maps of View-3 depth. For frame-based QoE measure, two SSIM index values are calculated: One between the uncompressed and the distorted texture frames of View-3, and the other between those of View-4. M is the average video distortion, calculated as the average of View-3 texture SSIM and View-4 texture SSIM values. Global Depth Combination of SSIM (GDC-SSIM) is computed as

$$GDC\text{-}SSIM = M\sqrt{Ddg} \qquad (4)$$

The final objective QoE measure is obtained by averaging Eq. 4 over all frames of the sequence.

VI. RESULTS & DISCUSSION

The subjective QoE measurements are conducted in three phases using the same stimuli, in order to compare three different subjective QoE models such as DSCQS, MSSP-DSCQS, SP-PC. The tests are performed according to ITU-R recommendation [14]. The assessors were not expert, having not much experience on auto-stereoscopic 3DTV. The assessors are allowed to participate only if they have passed tests for contrast sensitivity and far visual acuity, and random dot stereogram test. 14 assessors participated in the DSCQS test and 3 of them were detected as outlier and rejected. Similarly, 14 assessors participated in the MSSP-DSCQS test and 3 of them were rejected. On the other hand, 16 assessors participated in the SP-PC test and 1 of them was rejected.

A. SUBJECTIVE QoE MEASUREMENTS

The quality scores and confidence intervals which obtained from the three test results are gathered for each video and given in Table 2. Fig. 4 illustrates the quality scores and the confidence intervals. As shown in the figures, as the QP value increases the DMOS value also increases whereas the MOS and QS values decrease.

Table 2. *Subjective QoE scores and confidence intervals*

	Test	DMOS	MOS	QS
Balloons	QP35	11.3 ± 1.9	75.4 ± 1.6	70.9 + (-2.5, 8.7)
	QP40	21.5 ± 1.8	61.8 ± 1.8	32.2 + (-0.6, 1.9)
	QP45	44.8 ± 4.5	47.5 ± 2.2	3.9 + (-1.5, 1.6)
Kendo	QP35	14.4 ± 1.7	65.4 ± 2.3	64.8 + (-9.2, 2.4)
	QP40	29.2 ± 2.7	58.6 ± 2.0	30.7 + (-7.9, 2.3)
	QP45	47.5 ± 3.4	50.7 ± 2.3	8.0 + (-3.3, 4.8)
Undo Dancer	QP35	8.7 ± 2	77.1 ± 2.3	53.3 + (-3.9, 3.1)
	QP40	22.7 ± 2.4	76.4 ± 1.2	36.3 + (-2.6, 0.2)
	QP45	34.5 ± 3.5	69.6 ± 2.0	10.4 + (-3.6, 2.4)
Poznan Street	QP35	27.9 ± 4.0	72.5 ± 2.1	57.6 + (-6.3, 5.6)
	QP40	37.0 ± 3.3	72.1 ± 2.6	30.9 + (-2.0, 0.6)
	QP45	43.5 ± 4.4	68.9 ± 2.1	12.1 + (-1.7, 0.7)

In the DSCQS and MSSP-DSCQS tests, for Poznan Street and Undo Dancer difference in perceptual qualities could not be clearly perceived as for Kendo and Balloons. Because to preserve the original aspect ratio, Balloons and Kendo were cropped and fitted to the screen width whereas Undo Dancer and Poznan Street were down-sampled (958x540). Our observation is down-sampling spatially caused slight reduce in the sense of depth distortion. However, in the SP-PC test Poznan Street and Undo Dancer difference in perceptual qualities could be perceived easily since the methodology provides good discrimination for low QoE and high QoE cases allowing comparison of different test stimuli side-by-side. In the DSCQS and MSSP-DSCQS tests, test stimuli are always compared to reference stimuli.

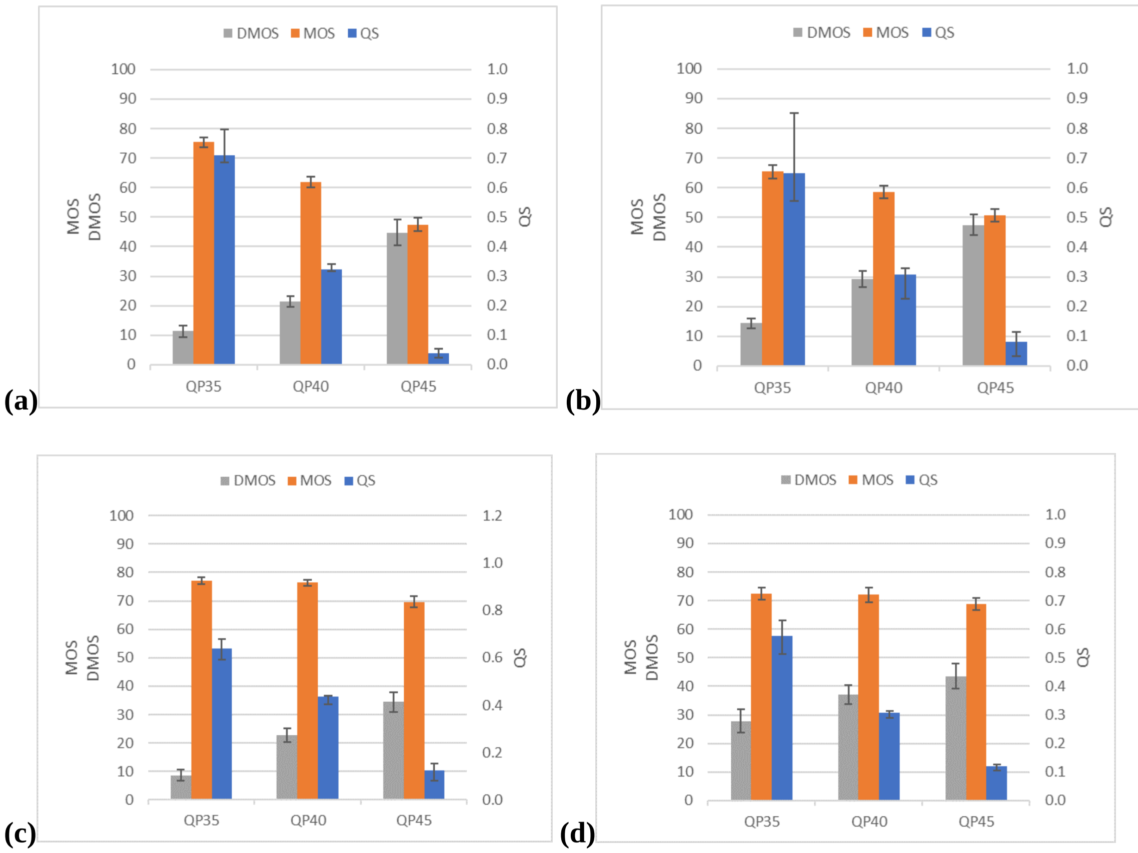


Figure 4. *Subjective QoE scores of* ***(a)*** *Balloons,* ***(b)*** *Kendo,* ***(c)*** *Undo Dancer,* ***(d)*** *Poznan Street*

B. CORRELATIONS WITH OBJECTIVE QoE

To demonstrate the performance of subjective QoE models, we computed the objective QoE values using the LDC-SSIM and GDC-SSIM metrics. Fig. 5 shows objective QoE curves versus the average bitrate for the MVD sequences. Average bitrate is calculated by dividing total bitrate by the number of views. The R-QoE curves show that as the QP value increases the average bitrate decreases and the QoE value decreases, as well. Namely, QP35 has a higher QoE value than that of QP40 and QP40 has a higher QoE value than that of QP45, for all sequences.

Table 3 and 4 show the correlations between the objective and subjective QoE measurements, evaluated by using two different methods such as Pearson’s Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC). For overall correlation computation Balloons and Kendo measurements are gathered as one group while Undo Dancer and Poznan Street measurements are gathered as another group. The CC values are calculated using LDC-SSIM and GDC-SSIM objective methods and are given in Table 3 and 4, respectively.

Correlation values obtained are very good and so close to each other for the single test video case (rows 1-4 in the tables), such as over 0.92 for PLCC and 1 for SROCC. However, for the overall computation case (rows 5-6 in the tables) the result turned out different. For “Balloons & Kendo”, the PLCC values are very close but the SROCC values of DSCQS are lower than those of MSSP- DSCQS and SP-PC. For “Dancer & Street”, it is observed that SP-PC is superior to MSSP-DSCQS and MSSP-DSCQS is better than DSCQS comparing both PLCC and SROCC values. Consequently, the best performance, the highest PLCC and SROCC values, is achieved with the SP-PC test and the LDC-SSIM metric.

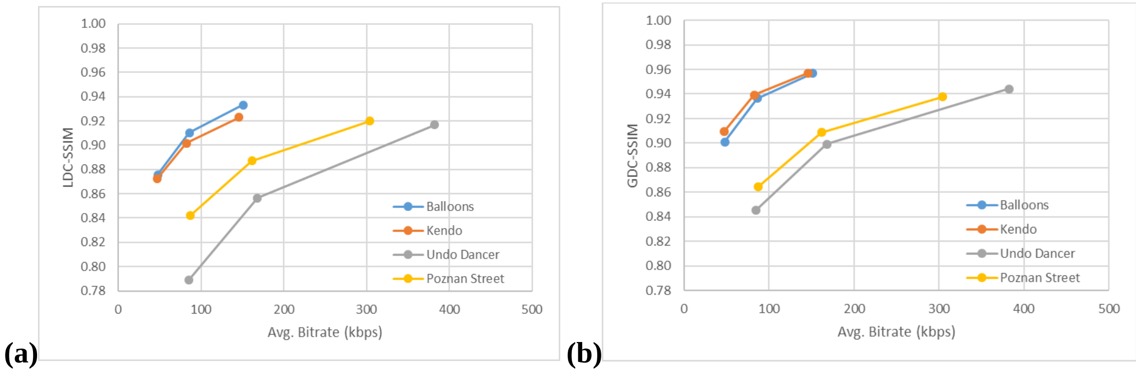


Figure 5. *R-QoE curves using* ***(a)*** *LDC-SSIM,* ***(b)*** *GDC-SSIM*

Table 3. *Correlation values (PLCC / SSROCC) between subjective and objective QoE measurements using the LDC-SSIM metric*

Test Video/s	DSCQS	MSSP- DSCQS	SP-PC
Balloons	-0.9947	0.9945	0.9781
Kendo	-0.9996	0.9989	0.9789
Undo Dancer	-0.9966	0.9189	0.9964
Poznan Street	-0.9824	0.9460	0.9818
Balloons & Kendo	-0.9961 / 0.6857	0.9687 / 0.9429	0.9682 / 0.9429
Undo Dancer & Poznan Street	-0.5316 / 0.4857	0.5836 / 0.6571	0.9085 / 0.9429

Table 4. *Correlation values (PLCC / SSROCC) between subjective and objective QoE measurements using the GDC-SSIM metric*

Test Video/s	DSCQS	MSSP- DSCQS	SP-PC
Balloons	-0.9976	0.9905	0.9707
Kendo	-0.9966	0.9949	0.9663
Undo Dancer	-0.9951	0.9253	0.9977
Poznan Street	-0.9764	0.9553	0.9757
Balloons & Kendo	-0.9641 / 0.6857	0.9312 / 0.8857	0.9626 / 0.8857
Undo Dancer & Poznan Street	-0.6975 / 0.6000	0.7276 / 0.7714	0.9693 / 0.8857

VII. Conclusion

We propose simple, reliable, and time-efficient MSSP-DSCQS and pair comparison (SP-PC) methods where the stimuli pair is presented simultaneously. We develop a subjective test setup for fair comparison of DSCQS, MSSP-DSCQS and SP-PC models. Experimental results show that the proposed models performed very good correlation with both LDC-SSIM and GDC-SSIM metrics. One advantage of the MSSP-DSCQS and SP-PC methods is that they bring time-efficiency by means of presenting stimuli simultaneously. Another advantage is higher accuracy which is proven by higher correlation values achieved with the test results when compared to those of the DSCQS test results. Furthermore, the SP-PC evaluation model is well-suited to assess small differences and discriminate similar QoE levels.

VIII. References

[1] *ITU-R Recommendations Methodology for the Subjective Assessment of the Quality of Television Pictures,* ITU-R Recommendation BT.500-11, 2002.

[2] *ITU-R Recommendations Methodology for the Subjective Assessment of Video Quality in Multimedia Applications*, ITU-R Recommendation BT.1788, 2007.

[3] N. Özbek, G. Ertan, and O. Karakuş, "Perceptual quality evaluation of asymmetric stereo video coding for efficient 3D rate scaling", *Turk. J. Elec. Eng. & Comp. Sci.*, vol. 22**,** no. 3, pp. 663–678, 2014.

[4] P. Aflaki, M. M. Hannuksela, and M. Gabbouj, "Subjective quality assessment of asymmetric stereoscopic 3D video," *Signal, Image and Video Processing*, vol. 9, pp. 1–15, 2015.

[5] J. Li, O. Kaller, F. D. Simone, J. Hakala, D. Juzska, and P. L. Callet, "Cross-lab Study on Preference of Experience in 3DTV: Influence from Display Technology and Test Environment", *IEEE Workshop: Quality of Multimedia Experience (QoMEX)*, 2013, pp. 46–47.

[6] *ITU-R Recommendations Subjective Video Quality Assessment Methods for Multimedia Applications*, ITU-R Recommendation P.910, 2008.

[7] J. Li, M. Barkowsky, and P. Le Callet, "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment", *IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 629–632.

[8] Z. Duanmu, A. Rehman, K. Zeng and Z. Wang, "Quality-of-Experience Prediction for Streaming Video," *IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 145–151.

[9] F. Qi, D. Zhao, X. Fan, and T. Jiang, "Stereoscopic video quality assessment based on visual attention and just-noticeable difference models," *Signal, Image and Video Processing*, vol. 10, no.4, pp. 737–744, 2016.

[10] E. Şenol and N. Özbek, "Quality of experience measurement of compressed multi-view video", *Signal Processing: Image Communication*, vol. 57, pp.147–159, 2017.

[11] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view Video plus Depth Representation and Coding", *IEEE International Conference on Image Processing (ICIP)*, 2007, pp. 201–204.

[12] C. Zhu, Y. Zhao, L. Yu, and M. Tanimoto, *3D-TV System with Depth-Image Based Rendering: Architectures Techniques and Challenges*, USA: Springer, 2013.

[13] G. J. Sullivan, J.-R. Ohm, J.-R. Han, and T. Wiegand, "Overview of Highy Efficiency Video Coding (HEVC) Standard", *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, no. 7, pp. 1649–1668, 2012.

[14] *ITU-R Recommendations Subjective Methods for the Assessment of Stereoscopic 3DTV Systems*, ITU-R Recommendation BT.2021, 2012.

[15] G.J. Sullivan, J. M. Boyce, Y. Chen, J.-R. Ohm, C. A. Segall, and A. Vetro, "Standardized Extensions of High Efficiency Video Coding (HEVC)", *IEEEJ. Sel. Top. Signal Proc.*, vol. 7, no. 6, pp. 1001–1016, 2013.

[16] L. Zhang, G. Tech, K. Wegner, and S. Yea, "3D-HEVC Test Model 5", *JCT-3V*, 2013.

[17] Fraunhofer Heinrich Hertz Institute, 3D-HEVC Reference Software. (2015) [Online]. Available: https://hevc.hhi.fraunhofer.de/3dhevc.

[18] D. Rusanovskyy, K. Müller, and A. Vetro, "Common Test Condition of 3DV Core Experiments", *JCT-3V*, 2012.

[19] J. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: application to scalable video coding", *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 882–893, 2011.

[20] J. Lee, L. Goldmann, and T. Ebrahimi, "Paired comparison-based subjective quality assessment of stereoscopic images", *Multimedia Tools and Applications*, vol. 67, no. 1, pp. 31–49, 2013.

[21] R. Bradley and M. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons", *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[22] F. Wickelmaier and C. Schmid, "A matlab function to estimate choice model parameters from paired comparison data", *Behavior Research Methods*, vol. 36, no. 1, pp. 29–40, 2004.

[23] V. De Silva, H. K. Arachchi, E. Ekmekcioglu, and A. Kondoz, "Toward an Impairment Metric for Stereoscopic Video: A Full-Reference Video Quality Metric to Assess Compressed Stereoscopic Video", *IEEE Transactions Image Processing*, vol. 22, no. 9, pp. 3392–3404, 2013.

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