Journal of Artificial Intelligence and Data Science (JAIDA)
e-ISSN: 2791-8335
https://dergipark.org.tr/pub/jaida

Vol. 2, No. 1, pp. 1-7, 2022
Research Article

Depth Perception Assessment of a 3D Video Based on Spatial Resolution

Gokce NUR YILMAZ[1,*], Yucel CIMTAY[1]

[1]TED University, Faculty of Engineering, Department of Computer Engineering, Turkey

Abstract

Burgeoning advances in 3 Dimensional (3D) video technologies can only be emphasized by considering the impact of these technologies on the perception of 3D videos from a user point of view. It is only possible to do this by considering the key factors that characterize the nature of a 3D video. Under the light of this fact, spatial resolution and perceptually significant depth levels, which are two effective factors for the depth perception of a 3D video, are used to develop a Reduced Reference (RR) model for the depth perception prediction of a 3D video. While determining the perceptually significant features, bilateral (abstraction) filter is exploited in this study. Structural SIMilarity metric (SSIM) is used to predict the depth perception enabled considering the degradation in the perceptually important features of depth maps having different spatial resolutions. The performance results of the developed model prove that it is quite effective in the depth perception prediction of a 3D video.

Keywords: 3D video; depth perception prediction; spatial resolution.

1. Introduction

Recent advances in 3 Dimensional (3D) video technologies enable 3D video multimedia services to provide better viewing experience to heterogeneous end-users. 3D video coding, transmission, adaptation are the frequently used names to these 3D video technologies. This viewing experience can be improved by investigating the depth perception assessment results of the end-users as feedback to modify the 3D content and service parameters. In order to make these parameters more reliable, the depth perception should be predicted as the most reliable way as possible [1]-[4].

In the current situation, the depth perception can be reliably assessed using subjective evaluation experiments. These experiments use human observers while assessing the depth perception of a 3D video. Therefore, they are considered quite effective for the depth perception assessment. However, they also cause costs in terms of money and time. There is also another way of assessing the depth perception of a 3D video in the literature namely objective model based assessment. The objective models can enable evaluation results in iterative and robust ways. Full Reference (FR), Reduced Reference (RR) and No Reference (NR) [1]-[3] are three objective model types. In the literature, there are objective model types developed for the depth perception assessment. The objective models developed for the video quality assessment of 2 Dimensional (2D) videos ((i.e., Peak-Signal-to-Noise-Ratio (PSNR) [5], Structural SIMilarity (SSIM) [6] and Video Quality Metric (VQM) [7]) are integrated with the results of the subjective experiments conducted for the depth perception evaluation in [8]. In [9], in order to develop a FR objective model for the depth perception assessment of a 3D video, first of all, subjective experiments are carried out to evaluate the sensitivity of the Human Visual System (HVS) towards binocular disparity, relative size, retinal blur is evaluated using the subjective experiments. Following this, a FR metric is developed for measuring the depth perception of a 3D video. When considering an FR assessment model, both the original and compressed videos are required at the end-user side. This can cause operational difficulties for the depth perception assessment. Therefore, the NR and RR metric types can be used alternatively to the FR metric type for assessing the depth perception of a 3D video. The depth range, vertical misalignment, and temporal consistency parameters are used to propose a NR depth perception model in [10]. In [11], the binocular disparity, region of depth relevance, frame based feature extraction and temporal information are exploited for the depth perception prediction in terms of a NR manner. In [3], a NR depth perception metric integrating binocular parallax, lateral motion, and aerial perspective cues for predicting perceived depth from the view of users is proposed. In [12], a RR model is proposed for the depth perception by extending the SSIM using edge information. It is a fact that a NR model does not need an original video during the assessment process.

Nevertheless, while evaluating the depth perception in the RR metric type, extra information extracted from the original and/or compressed 3D videos are needed. Thus, it can be stated that the RR models can provide more robust 3D video quality assessment than the NR metrics. However, in the literature, there is a lack of a reliable, efficient, and effective RR depth perception prediction model. Considering this, a RR model is developed in this study for predicting the depth perception of users towards a 3D video. Color plus depth map 3D video

[*] Corresponding author e-mail address: gokce.yilmaz@tedu.edu.tr

Received: 08/02/2022; Accepted: 21/02/2022

representation form enabled by a color video and its depth map counterpart has advantages over the other representation forms (e.g., left and right 3D video representation form) in terms of the network usage [1]-[3]. Therefore, the color plus depth map representation form is used while developing the proposed RR model in this study.

While trying to enhance the quality of the depth perception prediction results, key factors contributing the depth perception prediction should also be considered. In the light of this fact, spatial resolution and significant depth level features, which are key factors characterizing the nature of the depth perception [2], are used in the model development process. The significant depth level is a key factor due to the fact that it emphasizes perceptually important features (e.g., edges, shadows, etc.) for the depth perception. Therefore, in order to predict the depth perception of a 3D video, information degradation in the perceptually important features of the depth maps can be quantified. The reason why the spatial resolution is a key feature for the depth perception is that when the spatial resolution increases, the depth perception also improves. The SSIM is exploited to measure the depth perception considering the abstract (i.e., bilateral filtered) original and compressed depth map sequences at different spatial resolutions in this study.

The rest of the paper is organized as follows. The color plus depth map 3D video representation is introduced in Section 2. In Section 3, the proposed model is discussed. The performance assessment results and discussions are presented in Section 4. Finally, Section 5 concludes this study and points to the future work.

2. Color Plus Depth Map 3D Video Representation

Figure 1 [1]-[3] presents a snapshot captured using a 3D depth-range camera generating color plus depth map representation form for a 3D video. As can be observed from the snapshot, the spatial and temporal resolutions of the depth maps are similar to those of their related color images. All of the depth pixels have a related pixel in the color image. These pixels point to the distances of the related color image pixels to the observers. They take grey values from 0 to 255. 0 represents the furthest away pixel while 255 represents the closest pixel to the observers.

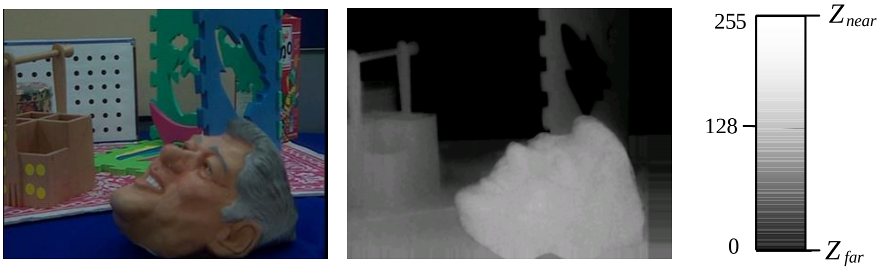


Figure 1. The *Orbi* sequence: (a) Color image (b) Related depth map

In order to render left and right views using the color images and depth maps, Depth-Image-Based Rendering (DIBR) is used. If a color and a depth map exist at the receiver side, a new image can be created by shifting the viewpoint on the screen. Thus, screen parallax is obtained on the screen. As a result of the parallax, different images for the left and right eyes of a viewer are obtained. These images give the depth effect. The basic principle of the DIBR technique relies on calculating screen parallax, which relies on pixel shifting [2]. In order to encode color plus depth stereoscopic video, the conventional 2D encoding techniques can be used [1]-[4].

3. Proposed Model

The proposed metric relies on the framework presented in Figure 2. As can be observed from the figure, the first step of developing the proposed metric is to abstract the original and compressed depth map sequences having different spatial resolutions. Quarter Common Intermediate Format (QCIF: 176×144 pixels), Common Intermediate Format (CIF: 352×288 pixels), and Standard Definition (SD: 704×576 pixels) are exploited as spatial resolutions in this study. In order abstract the depth map sequences at these spatial resolutions, the low contrast regions are simplified whereas the high contrast regions are emphasized using bilateral (abstraction) filter. In this way, the information meaningful for the depth perception assessment of the 3D video (e.g., the edges and shadows) is emphasized [1]-[3]. The bilateral filter utilizes Gaussian filtering method to replace each pixel value of an image with a weighted average of neighborhood pixel values. In this way, the image is smoothened using edge-preserving and noise reduction [12]. The bilateral filter is computed for a pixel using the equation below [1]:

$$F_a = \sum_{b\in\beta} w_{ab}(I_b / \sum_{b\in\beta} w_{ab}) \quad (1)$$

where, a is a pixel, I_b is at the intensity of pixel b in the kernel neighborhood β. The weighting coefficient at pixel b is computed as follows [2]:

$$w_{ab} = c(a,b)s(a,b) \quad (2)$$

where, *c*(*a*,*b*) and *s*(*a*,*b*) are clossiness and similarity kernel filters, respectively. These kernel filters are calculated as follows [3-4]:

$$c(a,b) = e^{(\frac{-1}{2}(a-b)^2/\sigma_c{}^2)} \quad (3)$$

$$c(a,b) = e^{(\frac{-1}{2}(a-b)^2/\sigma_c{}^2)} \quad (4)$$

Figure 3 presents the snapshots of the abstracted Breakdance and Chess depth map sequences. As seen from the figure, the perceptually important features in the abstracted depth map sequences are perceived better.

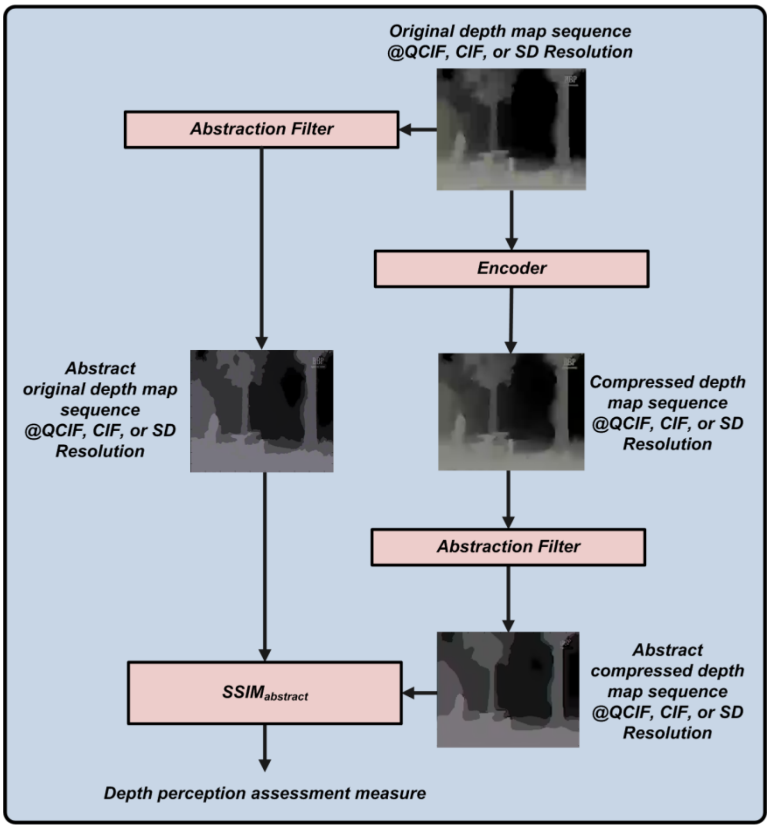
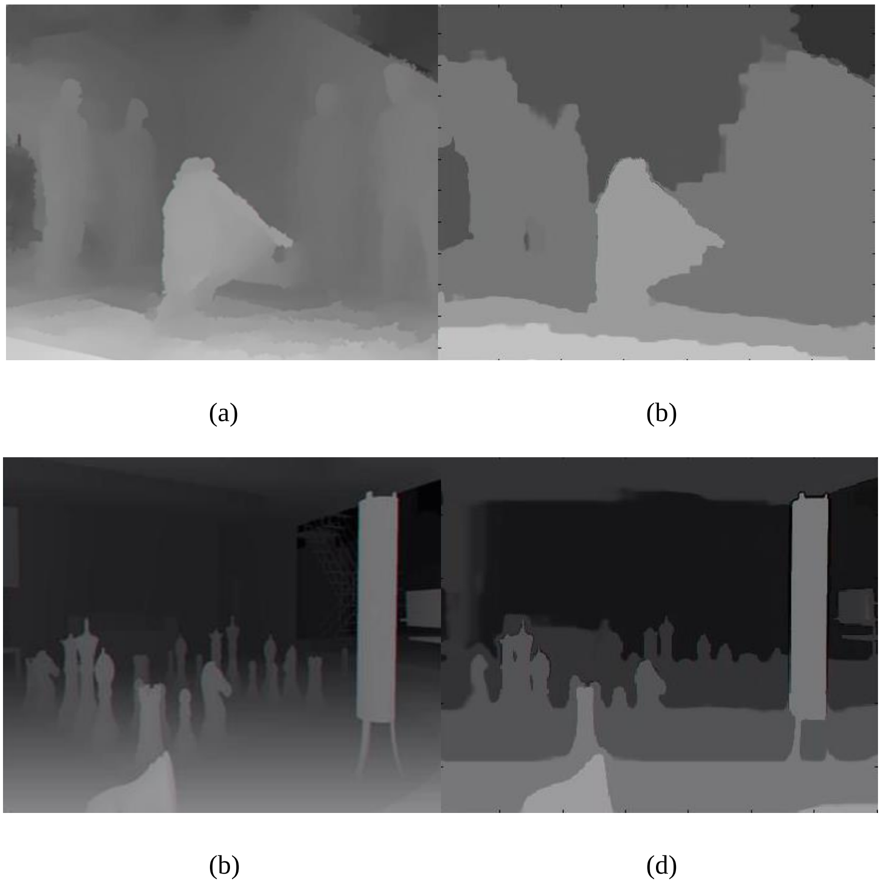

Original depth map sequence
@QCIF, CIF, or SD Resolution
Abstraction Filter
Encoder
Abstract original depth map sequence @QCIF, CIF, or SD Resolution
Compressed depth map sequence @QCIF, CIF, or SD Resolution
Abstraction Filter
SSIM$_{abstract}$
Abstract compressed depth map sequence @QCIF, CIF, or SD Resolution
Depth perception assessment measure

Figure 2. The framework of the proposed model



(a) (b)

(b) (d)

Figure 3. Abstracted Breakdance and Chess depth map sequences

As can also be seen from Figure 2, the last step of the proposed metric development is to utilize the SSIM to predict the performance of the proposed RR metric by comparing the abstracted information between the original and compressed depth map sequences at different spatial resolutions. $\text{SSIM}_{\text{abstract}}$ in Figure 2 refers to the SSIM measurement for the side information (i.e., the abstract original depth map sequence at QCIF, CIF, or SD spatial resolution) in the RR metric. The SSIM exploits the structural distortion of a distorted video compared to the original one. Thus, it is efficient to measure the degradation in the perceptually important features.

4. Results and Discussions

Three original 3D videos namely: Ice, Eagle, and Chess are used to derive the performance evaluation results. The snapshots of these 3D video sequences are illustrated in Figure 4. The original 3D video sequences are scaled down from their original resolutions (i.e., 1024×768 pixels) to QCIF, CIF, and SD resolutions using the Advanced Video Coding (AVC) 4-tap half-sample interpolation filter [13]. Four different bit rates (i.e., 512, 768, 1024, and 1536 kbps) are used to encode these down-scaled sequences with the Joint Scalable Video Model (JSVM) reference software version 9.13.1 at 25 fps [14]. 20% of the target bit rate is used to encode the depth map sequences whereas the 80% of the total bit rate is allocated for the color sequences of the 3D videos.

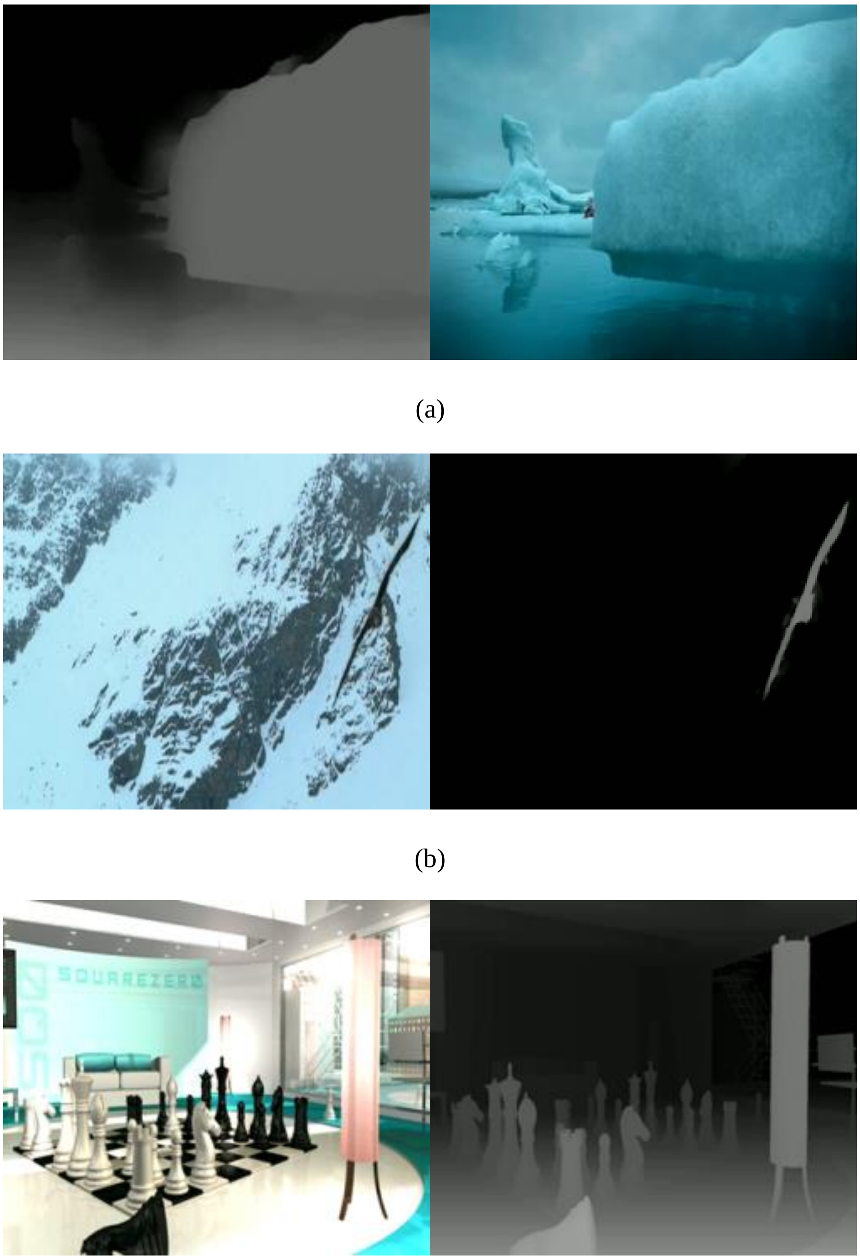
Then, subjective experiments are conducted using these encoded 3D video sequences. Double Stimulus Impairment Scale (DSIS) method is exploited during the experiments as suggested in International Telecommunication Union-Recommendation (ITU-R) BT-500.13 standard [15]. The 3D video sequences are presented in pairs: the first video is the original and the second video is the compressed one in this method. The participants are asked to evaluate the depth perception by comparing the impaired 3D videos with the original ones. During the experiments, an evaluation scale ranging from 1 to 5, which represents the lowest and highest depth perception respectively, is exploited for rating the 3D video sequences.

A 42" Philips multi-view auto-stereoscopic display, which has a resolution of 1920 × 1080 pixels, is utilized while displaying the 3D video sequences. 18 viewers (7 females and 11 males) volunteered in the experiments. Their ages range from 19 to 37 and they are all non-expert viewers. The outliers are omitted after obtaining the experiment results. Then, the Mean Opinion Scores (MOSs) and confidence intervals are computed exploiting 16 volunteers.

The VQM and SSIM$_{\text{abstract}}$ results of each 3D video sequence are computed for each bit rate (i.e., 512, 768, 1024, and 1536 kbps) at each spatial resolution (i.e., QCIF, CIF, and SD) considered in this study. The relationship between the MOS, VQM, and SSIM$_{\text{abstract}}$ results are approximated considering the symmetrical logistic function as suggested in ITU-R BT.500-13 [15]. The formulation of the symmetrical logistic function is shown in equation 5. [1][2]:

$$s = \frac{1}{1 + e^{(D-D_M)G}} \tag{5}$$

where, s is the normalized opinion score, D is the distortion parameter, and DM and G are constants.



(a)

(b)

(c)

Figure 4. Color texture and associated depth map of the (a) Ice (b) Eagle (c) Chess 3D videos

Considering the results of the symmetrical logistic function, Correlation Coefficient (CC), Route Mean Squared Error (RMSE), and Sum of Squares due to Error (SSE) metrics are computed to compare the performances of the MOSs, VQM, and SSIM$_{\text{abstract}}$. The CC, RMSE, and SS present the strength of the correlation between two variables, differences between the predicted values, and how well the correlation is calculated, respectively. The CC, RMSE, and SSE take values between 0 and 1. CC=1, RMSE=0, and SSE=0 present perfect results. The CC, RMSE, and SSE results of each 3D video sequence at each encoded bit rate (i.e., 512, 768, 1024, and 1536 kbps) are averaged for each spatial resolution (i.e., QCIF, CIF, and SD) as illustrated in Table 1.

As can be observed from the CC, RMSE, and SSE results in Table 1, the proposed metric perform better than those of the VQM for each of the spatial resolution. For instance, for the Ice video, VQM for the QCIF take 0.851, 0.07, and 0.175 CC, RMSE, and SSE values, respectively. However, the proposed metric enable a CC result of 0.903, a RMSE result of 0.041, and a SSE result of 0.147 for the QCIF spatial resolution of the Ice video. These outperforming results can also be observed for the rest of the spatial resolutions and videos.

Moreover, as can also be observed from Table 1, the proposed metric's performance gets better when the spatial resolution increases. In other words, the higher the spatial resolution of the depth map is, the better the abstraction result is. The reason behind this is that as discussed in Section 2 related to Figure 3, when the spatial resolution of the depth map sequences enhances, the visualization of the perceptually significant features also increases. Thus, it can be envisaged that the high spatial resolution abstracted depth maps particularly enhance the feeling of immersion into the scenes, which in turn affect the depth perception. These observations present the effectiveness of the proposed RR metric for assessing the depth perception of the 3D video.

Table 1. The performance evaluation results of the proposed model

3D Video	Depth Perception Metric At Different Spatial Resolutions	Depth Perception Results at Different Spatial Resolutions		
		CC	RMSE	SSE
Ice	VQM for QCIF	0.851	0.071	0.175
	VQM for CIF	0.869	0.063	0.169
	VQM for SD	0.883	0.056	0.162
	Proposed Metric for QCIF	0.903	0.041	0.147
	Proposed Metric for CIF	0.912	0.035	0.139
	Proposed Metric for SD	0.934	0.028	0.131
Eagle	VQM for QCIF	0.849	0.083	0.186
	VQM for CIF	0.864	0.074	0.177
	VQM for SD	0.871	0.067	0.171
	Proposed Metric for QCIF	0.908	0.049	0.153
	Proposed Metric for CIF	0.919	0.042	0.144
	Proposed Metric for SD	0.926	0.037	0.138
Chess	VQM for QCIF	0.843	0.088	0.193
	VQM for CIF	0.859	0.085	0.189
	VQM for SD	0.866	0.073	0.182
	Proposed Metric for QCIF	0.901	0.057	0.157
	Proposed Metric for CIF	0.909	0.051	0.149
	Proposed Metric for SD	0.917	0.044	0.143

5. Conclusions

In this paper, an RR metric has been proposed to predict the depth perception of the 3D video sequences. The color plus depth-map representation of the 3D video has been exploited in the proposed RR metric. The proposed RR metric considers the perceptually important depth levels and spatial resolution, which are key factors contributing to the nature of the 3D video, to measure the depth perception in a reliable, efficient, and effective way. The performance evaluation results of the proposed metric prove this fact. It has been envisaged that the development of the 3D video technologies in the consumer electronics market can be accelerated by exploiting the proposed metric. In our future study, the proposed metric will be integrated with a color video perception metric. In this way, both the video quality and depth perception parts of a 3D video perception will be considered.

Declaration of Interest

The authors declare that there is no conflict of interest.

Acknowledgements

This work has been supported by the Scientific and Technological Research Council of Turkey, Project Number: 114E551.

References

[1] G. Nur Yilmaz and F. Battisti, “Depth Perception Prediction of 3D Video for Ensuring Advanced Multimedia Services,” IEEE 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, Stockholm-Helsinki, Sweden-Finland, 3-5 June 2018.

[2] G. Nur and G. Bozdagi Akar, “An Abstraction Based Reduced Reference Depth Perception Metric for 3D Video,” International Conference on Image Processing (ICIP), Orlando, Florida, USA, 30 September-3 October 2012.

[3] G. Nur Yilmaz, “A Novel Depth Perception Prediction Metric for Advanced Multimedia Applications,” Springer Multimedia Systems, 2019.

[4] Perkis et.al., “QUALINET White Paper on Definitions of Immersive Media Experience (IMEx),” arXiv:2007.07032, 2020.

[5] Huynh-Thu and Ghanbari, “Scope of validity of PSNR in image/video quality M. assessment,” IET Electronics Letters, vol. 44, no. 13, pp. 800–801, Jun. 2008.

[6] Z. Wang, L. Lu, and A. C. Bovik, “Video Quality Assessment Based on Structural Distortion Measurement,” Proc. of Signal Processing: Image Com., vol. 19, no. 2, pp. 121-132, Feb. 2004.

[7] M.H. Pinson and S. Wolf, “A New Standardized Method for Objectively Measuring Video Quality,” IEEE Trans. Broadcasting, vol. 50, no. 3, pp. 312-322, Sep. 2004.

[8] D. Kim, D.Min, J. Oh, S. Jeon, and K. Sohn, “Depth Map Quality Metric for Three-Dimensional Video,” SPIE Stereoscopic Displays and Applications, San Jose, CA, USA, 18 Jan. 2009.

[9] D.V.S.X De Silva., G. Nur, E. Ekmekcioglu, and A. Kondoz, “QoE of 3D Media Delivery Systems,” Media Networks: Architectures, Applications, and Standards, CRC Press Taylor and Francis Group, May 2012.

[10] P. Lebreton, A. Raake, M. Barkowsky, P. Le Callet, “Evaluating Depth Perception of 3D Stereoscopic Videos,” IEEE Journal of Selected Topics in Signal Processing, vol.6, pp. 710-720, October 2012.

[11] T.E.R. Chaminda and M. G. Martini, “Quality Evaluation for Real-Time Video Services,” IEEE international Conference on Multimedia and Expo, 11-15 July 2011.

[12] G. Nur, S. Dogan, H. Kodikara Arachchi, and A.M. Kondoz, “Impact of Depth Map Spatial Resolution on 3D Video Quality and Depth Perception,” IEEE 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, Tampere, Finland, 7-9 June 2010.

[13] Advanced Video Coding for Generic Audiovisual Services, ITU-T Rec. H.264 (03/2005) Std., MPEG-4 AVC/H.264 Video Group.

[14] JSVM 9.13.1. CVS Server [Online]. Available Telnet: garcon.ient.rwth aachen.de:/cvs/jvt

[15] ITU-R BT.500–11, Methodology for the subjective assessment of the quality of television pictures.