EFFECTIVENESS OF 3VQM IN CAPTURING DEPTH INCONSISTENCIES

Dogancan Temel and Ghassan AlRegib

School of Electrical and Computer Engineering, Georgia Institute of Technology
Atlanta, GA, 30332-0250 USA
{cantemel, alregib} @ gatech.edu

ABSTRACT

The 3D video quality metric (3VQM) was proposed to evaluate the temporal and spatial variation of the depth errors for the depth values that would lead to inconsistencies between left and right views, fast changing disparities, and geometric distortions. Previously, we evaluated 3VQM against subjective scores. In this paper, we show the effectiveness of 3VQM in capturing errors and inconsistencies that exist in the rendered depth-based 3D videos. We further investigate how 3VQM could measure excessive disparities, fast changing disparities, geometric distortions, temporal flickering and/or spatial noise in the form of depth cues inconsistency. Results show that 3VQM best captures the depth inconsistencies based on errors in the reference views. However, the metric is not sensitive to depth map mild errors such as those resulting from blur. We also performed a subjective quality test and showed that 3VQM performs better than PSNR, weighted PSNR and SSIM in terms of accuracy, coherency and consistency.

Index Terms— Quality Assessment, Stereoscopic-3D, Depth Image Based Rendering

1. INTRODUCTION

Compared to 2D video, Three-Dimensional Television (3DTV) and Free Viewpoint Video (FVV) provide a realistic experience to the user by simulating the depth perception. Furthermore, FVV provides an interactive experience by enabling the user to navigate through the scene. In order to support depth perception, the overall system of 3D from content generation to display differs from existing 2D standards. In stereoscopic 3D (S3D) systems, we need both right and left views corresponding to a specific viewpoint. However, it is not feasible and not always possible to locate stereo camera systems at every single point that we would like to capture the scene. Moreover, as the number of viewpoints increase, 3D system will require higher computational capabilities and more complex coding and streaming techniques. To overcome feasibility issues and practical limits, researchers and developers have worked on various techniques such as depth image-based rendering (DIBR) [1].

Using DIBR, we need a single reference view and the corresponding depth map to synthesize a virtual view at new viewpoints. From content generation to the display side, each of the steps in the DIBR-based 3DTV processing chain affect the perceived quality. The technical report in [2] describes and categorizes stereoscopic artifacts that can occur inside a 3DTV processing chain. Performance of 3DTV and FVV systems are tested based on these artifacts in the literature. Compression and transmission contribute to the artifacts as outlined in [3] and [4]. It is established in the literature that the quality assessment of 3D videos inherently differs from quality assessment of 2D. As a result, researchers worked on 3D-specific concepts such as *naturalness* and *viewing experience* under varying blur and depth levels [5]. A broader discussion about challenges and advances in multimedia quality assessment can be found in [6].

In this paper, we specifically discuss the effectiveness of 3VQM in capturing certain types of distortions. The distortions we consider in this paper are limited to blur, compression artifacts, transmission losses and depth map estimation errors. We start by summarizing 3VQM in Section 2. Then, we describe the distortion types and analyze the performance of 3VQM in Section 3. We focus on validation in Section 4 and conclude the paper in Section 5.

2. A 3D VIDEO QUALITY MEASURE (3VQM)

3VQM is a 3D Video Quality Measure that objectifies the visual discomfort in the stereoscopic videos. We obtain *3VQM* by combining three distortion measures defined as spatial outliers (*SO*), temporal outliers (*TO*) and temporal inconsistencies (*TI*). We will briefly describe these distortion metrics but readers are encouraged to look at [7] for a detailed description.

Depth maps may not be accurate because of errors in depth estimation, rounding, compression and transmission. Therefore, we need to define an ideal depth map that would generate a visual distortion-free 3D video using DIBR. This definition implies that the video is free from DIBR-induced excessive disparities, fast changing disparities, geometric distortions, temporal flickering or spatial noise in the form of depth cues inconsistencies. Ideal depth is a function of the

color video for the view to be interpolated and it is used as a baseline to measure the errors in depth maps. Ideal depth can be estimated from the rendered virtual view intensity vector $\bar{I}_v$, the distortion-free view intensity vector $\bar{I}_o$, the received depth map $\bar{Z}$ vector, focal length F_v, relative location of the rendered view s ($+1$ for right and -1 for left), scaling factor α and the baseline B as follows:

$$\bar{Z}_{IDEAL} \approx \frac{sF_vB}{\alpha(\bar{I}_o - \bar{I}_v) + s\frac{F_vB}{\bar{Z}}} \quad (1)$$

We define $\Delta\mathbf{Z}$ as the difference between the *ideal* depth and *received* depth. Since we defined $\Delta\mathbf{Z}$, distortion metrics can be formulated as follows:

- *Spatial Outliers (SO):* Non-zero values of $\Delta\mathbf{Z}$ with non-uniform distribution results in relocation of pixels during the wrapping process. As a consequence, visual effects of these errors are spatially noticeable. Therefore, *SO* is a function of $\Delta\mathbf{Z}$ and can be expressed as the standard deviation of depth map errors.

$$\mathbf{SO} = STD(\Delta\mathbf{Z}) \quad (2)$$

- *Temporal Outliers:* Temporal variation of depth map errors leads to visual distortions that can appear as impulsive intensity changes around textured region and flickering around flat regions. To take into account these temporal variations, we can express *TO* as standard deviation of two depth map errors in time domain.

$$\mathbf{TO} = STD(\Delta\mathbf{Z}_K - \Delta\mathbf{Z}_{K-1}) \quad (3)$$

- *Temporal Inconsistencies:* Excessive and fast changing disparities result in visual distortions which can be modeled as the standard deviation of the difference of two depth values at different time instances.

$$\mathbf{TI} = STD(\mathbf{Z}_K - \mathbf{Z}_{K-1}) \quad (4)$$

We combine these distortion measures into a single 3D vision-based quality metric for *S3D* videos as follows:

$$\mathbf{3VQM} = K(1 - \mathbf{SO}(\mathbf{SO} \cap \mathbf{TO}))^a(1 - \mathbf{TI})^b(1 - \mathbf{TO})^c, \quad (5)$$

where K is a scale factor that we choose to be 5; and the constants a, b, and c are determined empirically. In [7], we suggested to use the following values: $a = 8$, $b = 8$ and $c = 6$.

3. PERFORMANCE EVALUATION OF 3VQM

In this section, we test the effectiveness of 3VQM in capturing errors and inconsistencies in the rendered depth-based 3D videos. At first, we will apply blur kernel under varying parameters to model the changes in naturalness and viewing experience as outlined in [5]. Then, we focus on compression artifacts and transmission losses that lead to visual distortions. We use *Balloons*, *Champagne Tower*, *Kendo*, *Lovebirds* and *Pantomime* sequences from *3DMobile* project. Virtual views are rendered using DIBR [1]. The hierarchical hole filling (HHF) is performed to avoid occlusion/disocclusion problems [8]. We also render virtual views using ground truth depth maps and reference views to get a baseline for comparison. PSNR and 3VQM results for the degraded video sets are given in Fig.1 and we will refer to this figure throughout this section.

3.1. Simulating Artifacts

We simulate the inaccuracy in depth maps using a Gaussian blur kernel. We implement different blur levels: 7×7 kernel with $\sigma = 2$, $\sigma = 5$, $\sigma = 10$ and 19×19 kernel with $\sigma = 10$, $\sigma = 20$ and $\sigma = 80$.

Compression artifacts lead to visual distortions that degrade the quality of user experience. We use *ver. 18.4* of *H.264/AVC* reference software to separately encode and decode ground truth depths and color videos [9]. We use *CABAC* as entropy coding method and perform different levels of compression with $QP = 28$, $QP = 40$ and $QP = 50$.

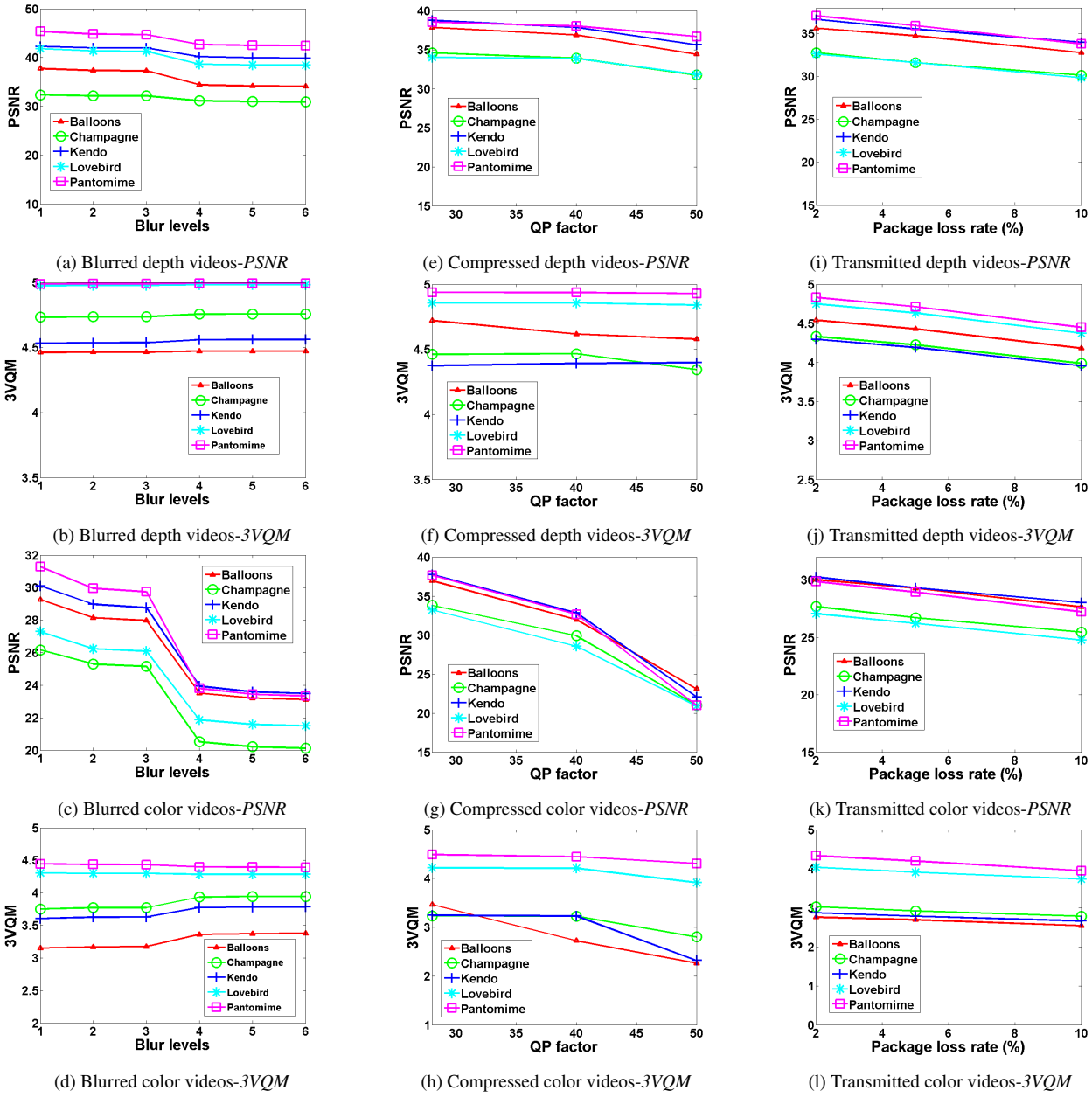
We packetize each frame as one packet and we use the *Gilbert Elliot* model to simulate the transmission losses. Usually, interpolation and error concealment algorithms are used to fill in the lost data and packets. In this work, we do not include any interpolation algorithm. *Packet loss rate* is set to *2%*, *5%* and *10%*. We perform different realizations of color and depth videos on five video sequences and using three different packet loss ratios.

3.2. Performance Evaluation

We synthesize the virtual views using reference color videos and degraded depth maps. Also we generate synthesized views using degraded color videos along with ground truth depth maps. We report the results where the distortion is applied to one channel, i.e., either the depth map or the reference video.

Compression and blurring of depth maps result in losing information that mostly corresponds to high frequency content or edges within the depth maps. But with depth maps having a simple structure and one color channel, compression and blurring lead to smoothing, which decreases the spatial and temporal variation. Hole filling also compensates for the inaccuracies that result from the smoothing of depth maps. The subjective quality of the rendered videos based on blurred and compressed depths are similar to the ones that are based on ground truth. PSNR decreases after a certain blur level, however 3VQM is almost insensitive to the blurring applied to the depth as it can be seen in parts *a*, *b*, *e* and *f* of Fig.1.

Color frames in the reference view are represented with three channels and the structure is inherently more complicated than depth. Objects located at the same distance with respect to the camera frame are represented with the same value in the depth maps. Whereas, pixel values of the same objects that are located at the same depth can significantly differ depending on the content in the color video. Thus, degradation in the color video directly degrades the rendered video.



(a) Blurred depth videos-*PSNR*

(e) Compressed depth videos-*PSNR*

(i) Transmitted depth videos-*PSNR*

(b) Blurred depth videos-*3VQM*

(f) Compressed depth videos-*3VQM*

(j) Transmitted depth videos-*3VQM*

(c) Blurred color videos-*PSNR*

(g) Compressed color videos-*PSNR*

(k) Transmitted color videos-*PSNR*

(d) Blurred color videos-*3VQM*

(h) Compressed color videos-*3VQM*

(l) Transmitted color videos-*3VQM*

Fig. 1. *3VQM* and *PSNR* results

Degraded color videos result in lower PSNR and 3VQM values for most of the sequences as it is given in parts *c*, *g*, and *h* of Fig. 1.

3VQM values decrease as we increase the level of compression applied to color videos as in Fig.1(h). In contrast, 3VQM gets slightly higher when we increase the level of blur as it is shown Fig.1(d). Since blurring smooths the image blockwise, the difference between the reference and rendered view becomes less in Eq.(1) and it results in higher *3VQM* because of lower $\Delta\mathbf{Z}$. Although 3VQM is more sensitive to compression and blur when it is applied to color video, the metric behaves differently with transmission losses. This is expected as with losses, the whole depth and color videos are lost and the metric value directly drops for all of the sequences as it is illustrated in Fig.1(i-l). We packetize depth and color videos that are already compressed with H.264. *QP* is set to 40 for all sequences. Transmission follows compression and errors are cumulative in the 3DTV processing chain. Thus, *PSNR* and *3VQM* decreases for all of the sequences after transmission.

The quantities SO and TO are functions of $\Delta\mathbf{Z}$. Therefore, 3VQM directly depends on the reliability of Z_{Ideal}, which is expressed in Eq.(1). All of the parameters except α are based on the DIBR configuration. However, we need to scale the difference of rendered and distortion-free views so that it will be effective in determining the value of Z_{Ideal}.

	RMSE	CC	ROCC	MAE	OR	σ_{DMOS}
Average PSNR	0.946	0.731	0.715	0.822	0.194	0.789
Weighted Average PSNR	0.935	0.755	0.777	0.790	0.194	0.789
Average SSIM	0.806	0.598	0.542	0.621	0.130	0.789
3VQM	0.616	0.894	0.789	0.517	0.000	1.008

Table 1: Validation scores for Average PSNR, Weighted Average PSNR, Average SSIM and 3VQM

If the difference term is not scaled with a reasonable value of α , the second term in the denominator will dominate the expression. Thus, Z_{Ideal} will be approximately equal to the received depth map and this results in *SO*=1 and *TO*=1. To scale the terms in the denominator into the same level, α is set to 120 for all of the sequences.

4. VALIDATION OF 3VQM

In addition to the videos from Mobile3D project, we captured stereo videos using Point Grey's BumbleBee2 camera. To simulate the degradation of quality, we perform H.264 based compression and estimated depth maps using stereo matching and 2D to 3D conversion methods instead of using ground truth depth map. We performed subjective quality assessment according to the requirements mentioned in [10]. Subjects evaluated the quality of video sequences with a discrete rating. Raw scores were collected and processed to give Difference Mean Opinion Scores (*DMOS*). 21 video sequences of 30 seconds length with both reference and distorted videos were used in the subjective test. Performance of *3VQM* is compared with PSNR, weighted average PSNR [11] and structural similarity index *(SSIM)* [12]. A more comprehensive analysis of *3VQM* including full-reference and no-reference approaches was submitted as a journal publication.

Validation scores are selected according to the VQEG recommendations. We use Root Mean Squared Error (RMSE), Pearson Linear Correlation Coefficient (CC), Spearman Rank Order Correlation Coefficient (ROCC), Mean Absolute Error (MAE), Outlier Ratio (OR) and the standard deviation of the DMOS values (σ_{DMOS}).We define outliers as the points whose distance from the reference is greater than twice the *DMOS* standard deviation. High *CC* and *ROCC* shows coherency, low *RMSE* and *MAE* indicates accuracy and low *OR* represents consistency. As it is given in Table 1, 3VQM is the most accurate, coherent, and consistent among all objective measures represented in this paper.

5. CONCLUSION

We evaluated the effectiveness of *3VQM* in capturing depth inconsistencies by simulating compression artifacts, transmission losses and depth map estimation errors. According to the simulation results, *3VQM* captures the depth inconsistencies based on errors in the reference views more effectively than errors in the depth map. Errors based on smoothing are not considered as degradation since they lead to decrease in temporal and spatial variations. We performed subjective quality assessment to validate 3VQM and we showed that 3VQM is the most accurate, coherent, and consistent among all objective measures represented in this paper.

6. REFERENCES

[1] C. Fehn, "Depth-image-based Rendering (DIBR), Compression, And Transmission For A New Approach On 3DTV," *Proc. of SPIE*, vol. 5291, pp. 93–104, 2004.

[2] A. Boev, D. Hollosi, and A. Gotchev, "Classification of stereoscopic artefacts," *Technical report D5.1., MOBILE3DTV Project*, 2008.

[3] P. Joveluro, H. Malekmohamadi, W.A.C. Fernando, and A.M. Kondoz, "Perceptual video quality metric for 3d video quality assessment," in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2010.

[4] Y. Liu, j. Wang, and H. Zhang, "Depth image-based temporal error concealment for 3-d video transmission," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 4, pp. 600 –604, april 2010.

[5] R. G. Kaptein, A. Kuijsters, M. T. M. Lambooij, W. A. IJsselsteijn, and I. Heynderickx, "Performance evaluation of 3d-tv systems," 2008.

[6] F. Porikli, A. Bovik, C. Plack, G. AlRegib, J. Farrell, P. Le Callet, Quan Huynh-Thu, S. Moller, and S. Winkler, "Multimedia quality assessment [dsp forum]," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 164 –177, nov. 2011.

[7] M. Solh, G. AlRegib, and J. M. Bauza, "3VQM: A vision-based quality measure for DIBR-based 3D videos," in *2011 IEEE ICME*, July 2011, pp. 1 –6.

[8] M. Solh and G. AlRegib, "Hierarchical Hole-Filling (HHF): Depth Image Based Rendering without Depth Map Filtering for 3D-TV," in *IEEE MMSP'10*, Saint-Malo, France, 2010.

[9] Fraunhofer Heinrich Hertz Institut, "H.264/avc software coordination - jm 18.4," .

[10] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "New Requirements Of Subjective Video Quality Assessment-Methodologies For 3DTV," in *VPQM*, Scottsdale,US, 2010.

[11] N. Ozbek, A.Tekalp, and E. Tunali, "Rate Allocation Between Views in Scalable Stereo Video Coding using an Objective Stereo Video Quality Measure," in *ICASSP*, April 2007.

[12] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. on Image Proc.*, vol. 13, pp. 600–612, 2004.