

Quality Assessment of Synthesized 3D Video with Distorted Depth Map

Hsin-Che Liu and Hsueh-Ming Hang

Department of Electronics Engineering, National Chiao Tung University
Hsinchu, Taiwan
hmhang@mail.nctu.edu.tw

Abstract— In the virtual-view 3D video coding system, both the RGB image data and the depth maps are compressed and translated to the receivers. After compression, the depth maps are distorted and may cause visible artifacts on the synthesized video. We study the visual effect of compressed depth maps on the synthesized video and develop a quality assessment model that predicts the subjective quality. We use HEVC Test Model (HTM) to compress the depth maps. The distorted depth value may lead to ghost artifacts around object edges and unnatural object motion on the synthesized video. In our proposed quality assessment (QA) model, we use SSIM to compute the basic score of stereo image pair; we extract the edge, motion, and depth features of stereo pairs and combine them to form a local weight to increase the sensitivity of the noticeable regions. We use the binocular perception model to calculate the score of stereo pairs. We conduct our own subjective tests. The results of our experiments show that our model has a better match to the subjective scores when it is compared with the other existing metrics.

I. INTRODUCTION

The 3D perception is often made by viewing two different views in two eyes, and then they are combined by the Human Visual System (HVS). The ISO/IEC Moving Picture Expert Group (MPEG) is in the process of specifying the 3D video coding (3DVC) standards based on the multiple-view plus depth (MVD) format. It assumes the input is a 2-view (or more views) video, and each view has its corresponding depth map, which can be captured by depth sensors or generated by a depth estimation algorithm. These color and depth images are then compressed by a 3D video coder. At the receiver, the virtual view images are generated by a view synthesis algorithm. Either the transmitted views or the synthesized views and their mixtures can be displayed on a 3D monitor. With the popularity of 3D virtual view systems, how to predict the quality of the synthesized stereo images becomes an important issue.

In recent years, several research groups studied the 3D quality assessment topic and some of them provide their database to the public on the website. For different purposes of the 3D QA research, these databases can be classified into a few categories. For example, the focus of Benoit et al. [1] is on the color distortion. They compress images or videos by JPEG and JPEG2000 or blur images to observe their effect on the subjective scores. Lavoue et al. [2] work on the quality assessment for 3D computer graphics. Two types of distortion (noise addition and smoothing) were applied with different strengths and on 4 reference objects. Goldmann et al. [3]

capture the nature scenes with a static camera at different camera distances in the range 10-50 cm. Bosc et al. [4] use different synthesis algorithms to recreate images. There are seven synthesis algorithms on three sequences. In this paper, we interest in the effect of distorted depth map on the synthesis video. Because our target is different from the previous ones, we construct our own test database which consists of six scenes. We use the test videos provided by the ITU/MPEG standardization committee for specifying the Advanced Video Coding (AVC, H.264) and High Efficiency Video Coding (HEVC, H.265) 3D standards. The depth maps are compressed by HTM (HEVC Test Model-8.0) and use the original color images and compressed depth maps to synthesis the virtual view image/video. The synthesis software is VSRS (View Synthesis Reference software 3.5).

In this paper, we propose a new visual quality assessment (VQA) model based on our collected data to assess the visual quality of synthesized video, of which the depth map is distorted by compression. The metric uses the extracted features and their combination with different weighting, so that the artifacts can be properly addressed. We provide experimental results on the proposed method, and demonstrate that the proposed scheme can improve the correlation between the score of the computational QA model and the subjective scores.

In this paper, we first introduce briefly the depth coding principles and the artifacts caused by distorted depth maps in Section II. Section III is the proposed computational objective model. Section IV is the results of subjective experiments. Finally, Section V concludes this presentation.

II. DEPTH CODING AND ARTIFACTS CAUSED BY DEPTH ERROR

In the HTM depth coding process, there are three kinds of prediction models (intra-prediction, motion-compensated prediction and disparity-compensated prediction) [5]. The depth maps are quantified and divided into coding blocks with different sizes. Each block chooses the prediction model that has the least Rate Distortions cost (RD cost). After the quantization process, the coding blocks are divided into smaller blocks until the RD cost of the original block size is less than the sum of RD costs using the smaller blocks. The motion-compensated prediction and disparity-compensated prediction code the blocks by using the information of other coded frames. Unlike the color pixel coding, specific techniques have been developed for depth value coding. Limited by space, we only

describe briefly two intra coding modes. They are Planar Mode (1 segments) and DMM Mode 1 (2 segments). The purpose is to give readers some ideas of the sources of compression distortion.

A. Planar Mode

If the coding blocks are grouped into one segment, they are considered as the smooth regions. The Planar Mode saves only the four depth values at each corner and uses the corner information to interpolate the other depth values of each pixel in the block. In Fig. 1 (a), the Planar Mode saves the depth value of the four corners (0,0), (0,7), (7,0) and (7,7), and then interpolates the other values in the block.

B. Explicit Wedgelet Mode

In the edge regions, the block will be partitioned into two segments. The Wedgelet Mode saves the four values at the corners and the start and end point of the segmentation line (boundary). It then uses the segment mean value to represent all the pixels in one segment. The mean value of each segment is computed based on the mean depth value of corners belonging to it. Clearly, using the mean value to represent other values is imprecise. A Depth Lookup Table (DLT) is formed to compensate the residual values to match the original depth values. In Fig. 1 (b), the Wedgelet Mode saves the depth value of four corners (0,0), (0,7), (7,0), (7,7), and the start and end point of the segment line (1,7) and (7,3). The mean value of segment 1 (dark color) is the mean value of (0,0), (0,7) and (7,0). On the other hand, the mean value of segment 2 (light color) is the depth value of (7,7).

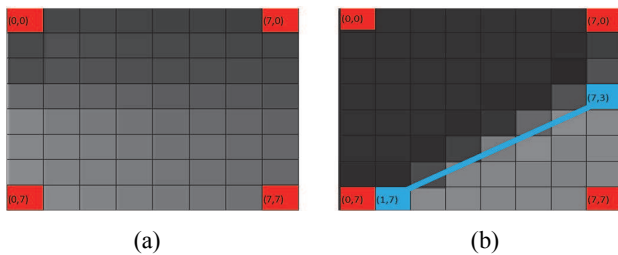


Fig.1 the example of (a) Planar mode (b) Wedgelet Mode

The incorrect depth value causes the shifting phenomenon in viewing, because the image may be warp to wrong positions. As shown in Fig. 2 (a), the P1 and P2 represent the projection paths of the object into the camera 1 and the camera 2, and the P is the projection path of the virtual camera. They all have the same depth values assuming all the cameras are in parallel. If the depth values associated with P1 and P2 have coding errors, then Fig. 2 (b) shows that object is closer to the virtual camera in the case of a smaller depth value. On the image plane, the object location x is changed to location x' . The difference between x and x' causes the shift artifact, as illustrated by Fig.

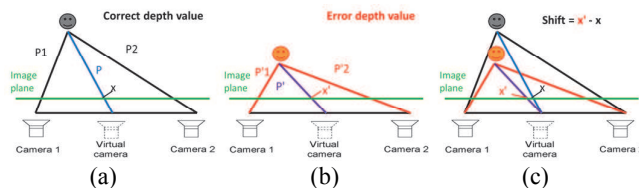


Fig. 2 (a) Correct depth; (b) Incorrect depth; (c) Combine (a) and (b).

2 (c). An example of this artifact on the synthesized image is showed in Fig. 3.

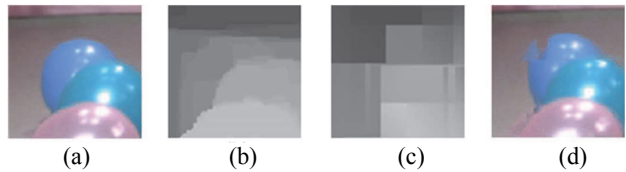


Fig. 3 (a) Reference image; (b) Reference depth map; (c) Distorted depth map; and (d) Synthesized image.

The shift artifact could result in the unnatural movement in a video. If the depth value of the object in the previous frame is different from that in the current frame. The object positions on the image plane are thus shifted. It looks like that the object moves forward or backward. This problem usually appears in the moving regions, because the foreground objects move into or out of the coding blocks, which causing the large changes of depth values at the four corners used in coding. As the block uses the Planar mode to coding, the edge of foreground and background is blurred. Therefore, the same object may have different depth values between frames.

III. COMPUTATIONAL QA MODEL

In our subjective experiments, we asked every viewer to identify the regions with annoying artifacts. We check the answered regions against the Structural Similarity (SSIM) metric map [6]. SSIM can easily detect the region of the shift artifacts. However, not all shift artifacts can be easily detected by human. An example is given in Fig. 4. (b) and (d) are the typical shift artifacts, which can easily be detected by human. Fig. 4(f) contains the shift artifacts on the road portion (Fig. 4(e)). Fig. 4(g) shows the SSIM scores. Although the SSIM detects the artifacts (dark regions), they are hard to be observed by human. These regions have heavily distorted depth maps. However, these regions are smooth, and the shift artifacts are less noticeable to the human. On the other hand, the SSIM is calculated pixel-by-pixel, and they are sensitive to object shifts. Thus, we use the edge information as one of our features.

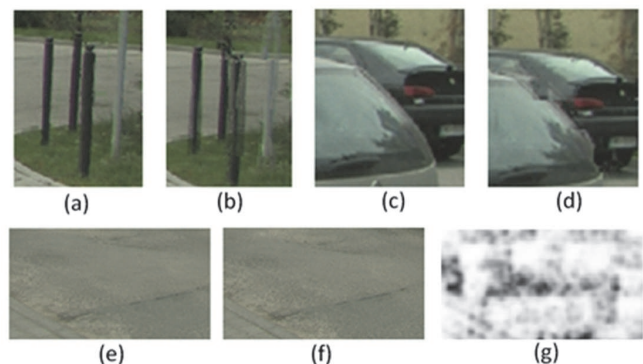


Fig. 4. (a) (b) (c) (d) are the examples of significant shift artifacts. (e) (f) is the example of less obvious shift artifact. (a), (c) and (e) are the reference images. (b), (d) and (f) are the synthesized images. (g) is the SSIM map between (e) and (f).

In addition to the above two cases, there are other cases that the obvious artifacts are less noticeable. For example, people pay less attention to the faraway background. Therefore, many 3D quality assessment models also consider the depth information as an important factor. Thus, our second feature is the depth information. The last feature of our model is motion. Because people usually pay attention to large moving objects and the unnatural movements easily get attention. We use these three features to compute the local weights of each local region.

Our proposed QA model is divided into two parts. The first part computes the SSIM of the stereo video, and the second part is generating weightings for each extracted local features of video. The proposed method computes the score of each frame and combines them into a representative score of the entire video. For each frame, we divide an image into 8-by-8 blocks, and the Structural Similarity (SSIM) metric and the feature extraction are performed inside these 8-by-8 blocks. We then combine the scores of right and left view into the score of a frame. The flow chart of our proposed model is shown below (Fig. 5).

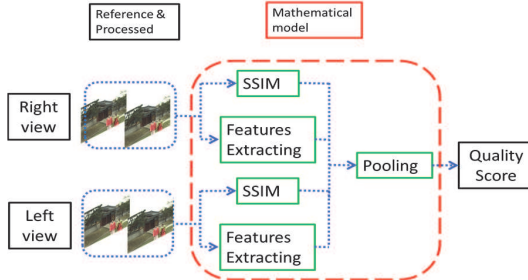


Fig. 5. The flow chart of the proposed model.

This SSIM metric was proposed by Wang, et al. [6]. The SSIM index consists of three components: luminance, contrast and structure.

$$L(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, C(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, S(x, y) = \frac{2\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (1)$$

$$SSIM(x, y) = [L(x, y)]^\alpha [C(x, y)]^\beta [S(x, y)]^\gamma, \quad (2)$$

where x and y are the reference and distorted image, respectively. The luminance, contrast, and structure can be computed by combining the means, standard deviations, and correlation coefficient of the images. Finally, the overall image quality is evaluated from the average SSIM.

The edge factor is extracting by the ‘‘Sobel’’ edge detector applied to the depth map. The map after edge detection is $edge(x, y)$. We define the number of edge as the factor of edge.

$$E(u, v) = \frac{1}{8 \times 8} \sum_{(x, y) \in block(u, v)} edge(x, y) \quad (3)$$

The (u, v) pair is the index of blocks in each frame, and (x, y) is the index of pixels in a block. Each $edge(x, y)$ is assigned with value 1, if this pixel (x, y) belongs to an edge. Otherwise, its value is 0.

The motion factor is extracted by a 4-level hierarchical block matching algorithm. Each level down-sample the test image by a factor of 2, and the search method is the four-step

search. The block size is 8-by-8. The motion vector map stores the motion vector magnitude, $motion(u, v)$ as defined below.

$$M(u, v) = \begin{cases} 0 & ; \text{if } motion(u, v) < motion_{thd} \\ \frac{1}{motion(u, v)} & ; \text{otherwise} \end{cases} \quad (4)$$

First, we classify the entire image into motion and non-motion regions. For each block, if the $motion(u, v)$ is less than the threshold, it is classified as non-motion, and the motion factor is 0. Second, we take into account the ghost and afterimage issues. In Fig. 4 (c)(d), the ghost artifact is easily detected if they are in the non-motion images. However, the ghost artifacts are similar to the afterimages in the video. It less noticed if the object speed gets higher. Thus, we define the motion factor as the reciprocal of the motion magnitude.

The depth information is generated by the depth estimation methods. We compute the disparity to estimate the perceptive depth value. Eq.(5) shows the relationship between disparity and perceptive depth.

$$Z_p = f \frac{T}{disparity} \quad (5)$$

The disparity gets bigger, the object becomes closer. So, the disparity factor $D(u, v)$ is defined to be the $disparity(u, v)$.

After extracting all feature factors, we combine three factors into a set of local weighting for each frame. The total weight is computed by (6).

$$w(u, v) = \alpha \times E(u, v) + \beta \times M(u, v) + \gamma \times D(u, v) \quad (6)$$

The definitions of α , β and γ are given below:

$$\alpha = \frac{1}{E_{max}}, \quad \beta = \frac{h/h_0}{M_{max} \times fr/fr_0}, \quad \gamma = \frac{h/h_0}{D_{max}} \quad (7)$$

We normalize all features by their maximum values separately to the range [0 1]. The motion estimation is pixel based, so the motion feature is affected by the resolution of sequence and the frame rate. To deal with this effect, we multiply the ratio of h and h_0 and fr_0 and fr , individually, where h is the picture height, and fr is the frame rate. In our test sequences, h_0 is 768, and fr_0 is 30. The disparity weight also needs to be adjusted by the sequence height. The final score of each block in i th frame is:

$$score_i(u, v) = \frac{w(u, v) \times ssim(u, v)}{\frac{1}{n} \sum_{(u, v) \in i \text{th frame}} w(u, v)} \quad (8)$$

To calculate the score of a stereo image pair, we incorporate the Binocular Perception Model [7] into my model. For this model, the subjective 3D image quality is determined by the mixture of the higher and lower quality images. The equation of Binocular Perception Model is as follows:

$$Q_{binocular} = \{w \cdot Q_{high}^n + (1 - w) \cdot Q_{low}^n\}^{\frac{1}{n}} \quad (9)$$

In our experiment, w is set to 0.86 and n is 1. We compare the score of each block in the right image with that of the corresponding left image block, and the bigger one is Q_{high} , and the other is Q_{low} . Compute all the scores of stereo block pairs in a frame and use the average of the 5% worst block scores in a frame to form the score of this frame. Finally, we compute the average of all frames in a sequence to form the final score of this stereo video.

V. EXPERIMENT RESULTS

In our experiments, we focus on the effect of distorted depth maps. We only compress the depth maps and use the original color images to synthesis the virtual images. To reduce the effect of synthesis algorithms, the reference videos are produced also by the same synthesis algorithm using the original depth maps.

TABLE I. THE SUBJECTIVE TEST SETUP.

| | |
|----------------------|---|
| Participants | 26 people (man 19/woman 7 /age: 20~25) |
| Methodology | Double Stimulus Continuous impairment Scale (DSIS) [8] |
| Sequence | Six multi-view plus depth sequences provided by MPEG 3DVC |
| Type | 5 levels of quantization parameters (QP=16, 27, 36, 43, 48) |
| Number of test video | 5*6=30 |

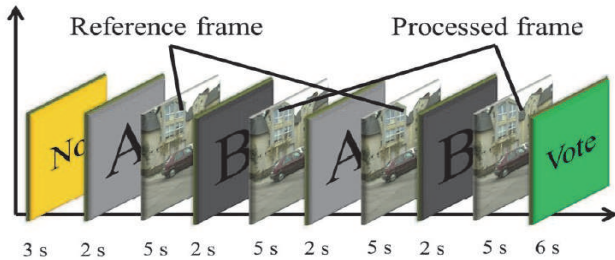


Fig. 6. The structure of DSIS

We first conduct the subjective video evaluation. The test set up is summarized in Table I. In our experiment, we add some dummy sequences to test the reliability of subjective scores. The dummy sequences are the repeating sequences or the reference to reference sequences. If some participants give different score to the same test video or give the lower score to the reference sequence to reference sequences, the entire data sets of these participants are dropped.

We compare the performance of our model with four metrics (PSNR, SSM, MSSIM, VIF(Visual Information Fidelity)). And we use the PLCC, SROCC and RMSE to evaluate the performance of all metrics. To remove the effect of nonlinear relationship on computing the correlation coefficient, Video Quality Experts Group (VQEG) Full Reference Television (FRTV) Phase II report [9] recommends a process that measures the performance of the objective QA metric and we follow this process.

TABLE II. THE QA MODEL COMPARISON.

| | PLCC | SROCC | RMSE |
|-----------------|---------------|---------------|---------------|
| PSNR | 0.7173 | 0.85 | 1.6006 |
| SSIM | 0.5956 | 0.7597 | 2.0173 |
| MSSIM | 0.5682 | 0.8008 | 2.0476 |
| VIF | 0.7067 | 0.7539 | 1.7080 |
| Proposed | 0.9280 | 0.8460 | 0.7440 |

As Table II indicates, our QA model is better than these existing models except that our SROCC values are slightly lower than that of PSNR.

VI. CONCLUSIONS

In this paper, we propose a computational quality assessment model to estimate the quality of distorted video synthesized by a distorted depth map. We observe some special phenomena that do not occur in the conventional 3D video not generated by a virtual view synthesizer. In our proposed model, we extract edge, motion and depth features to compute the local weighting and thus enhance the effect of the “noticeable” regions with visible artifacts. Overall, we propose a new 3D video quality metric. The experimental results indicate that the proposed method has a higher correlation with the subjective scores (higher PLCC and lower RMSE score).

VII. ACKNOWLEDGEMENT

This work was supported in part by the NSC, Taiwan under Grant NSC 103-2218-E-009-009, 101-2221-E-009-136-MY3 and by the Aim for the Top University Project of National Chiao Tung University, Taiwan.

IV. REFERENCES

- [1] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, “Quality assessment of stereoscopic images,” *EURASIP Journal on Image and Video Processing*, 2008.
- [2] G. Lavoué, E. Drelie Gelasca, F. Dupont, A. Baskurt, and T. Ebrahimi, “Perceptually driven 3D distance metrics with application to watermarking,” *In Proceeding of SPIE*, vol. 6312, 2006.
- [3] L. Goldmann, F. De Simone, and T. Ebrahimi, “Impact of acquisition distortion on the quality of stereoscopic images,” *5th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Scottsdale, USA, 2010.
- [4] E. Bosc, R. Pèpion, P. Le Callet, M. Köppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, “Towards a new quality metric for 3-D synthesized view assessment,” *IEEE journal on selected Topics in Signal Processing*, 2011.
- [5] L. Zhang, G. Tech, K. Wegner, and S. Yea, “3D-HEVC Test Model 5,” Joint Collaborative Team on 3D Video Coding Extensions (JCT-3V) document JCT3V-E1005, 5th Meeting: Vienna, AT, July – Aug. 2013.
- [6] 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, pp. 600–612, April 2004.
- [7] D.V. Meegan, L.B. Stelmach, and W.J. Tam, “Unequal weighting of monocular inputs in binocular combination: Implications for the compression of stereoscopic imagery,” *Journal of Experimental Psychology: Applied*, vol. 7, pp. 143, 2001.
- [8] ITU-R Recommendation BT.500-13, “Methodology for the subjective assessment of the quality of the television pictures”, ITU-R Telecommunication Standardization Bureau, 2012.
- [9] Video Quality Experts Group, “Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment,” VQEG, August 2003.