P-45: A Quality Measurement Based on Object Formation for 3D Contents

*Chung-Te Li, Yen-Chieh Lai, Chien Wu, Chao-Chung Cheng, and Liang-Gee Chen**

Graduate Institute of Electronics Engineering, National Taiwan University, Taipei City, Taiwan

Abstract

In this paper, we propose a quality metric for three-dimensional images or videos. The goal of this work is to provide a quality measurement from the relationship between the depth image quality and the three-dimensional perception. From our observation, we found that there is a relation between the perception of depth and the formation of object. The consistency for depth and object perceptions has a strong influence on the acceptable visual quality in three-dimensional images. Since the visual quality depends on the consistency, additional work is necessary for developing objective metrics based on this concept. In this paper, we propose a quality measurement metric for three-dimensional images by checking the consistency between depth and object. The measured visual quality can be used as feedback for developing depth generation algorithms, which is important for 3D industry.

1. Introduction

2D-to-3D conversion algorithms have been widely used in 3D display systems. In the past, most of the 2D-to-3D conversion algorithms have been proposed from physical concern and depth cues. With recent advances in 3D image or video processing, perceptual consideration has also been demonstrated for depth generation. Therefore, quality assessments of three-dimensional images or videos have become more and more important.

For conventional two-dimensional images and videos, researchers have paid great attention to image/video quality assessments in the decades. However, the quality assessment of 3D images and videos is still a challenging research task. Objective assessment is greatly demanded due to the huge time consuming and cost for subjective assessment. For three-dimensional images and videos, it is important to find suitable objective quality assessments due to health issues especially. For example, Nintendo unveils 3DS and follows up with a statement about the possible dangers to children. Samsung also issues a warning about potential health risk to certain viewers in 3D-TV. 3D-TV cannot be proved that there is no harm for human at all until now. Therefore, the objective assessments for 3D videos are eager for the 3D display systems.

In this paper, we try to propose a quality metric for three-dimensional images or videos based on the correspondence between depth and object. In order to prove our proposed effectiveness some existing 2D-to-3D conversion algorithms will be reviewed and tested both subjectively and objectively.

The paper is organized as follows: Section 2 reviews the backgrounds of 2D-to-3D conversion algorithms and 3D feeling. Section 3 provides the details of our proposed objective measurements. Section 4 gives the experiment results. Finally, Section 5 presents our conclusions.

2. Backgrounds and Previous Study

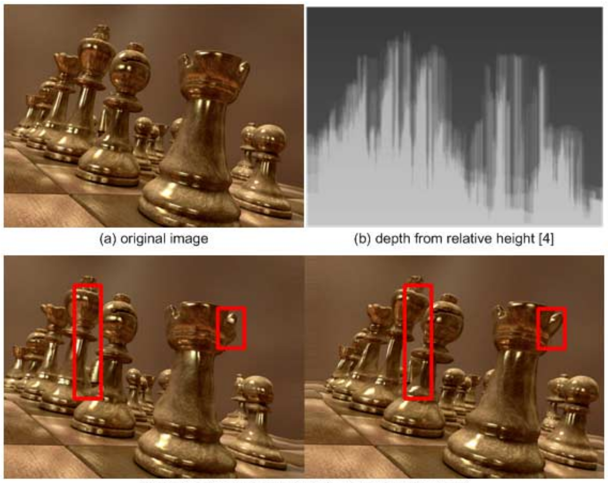
Existing methods of 2D-to-3D conversion can be classified into two types: pictorial-depth-cue methods, and motion-based methods. Pictorial-depth-cue methods, which include depth from edge, depth from relative height, depth from texture gradient, depth from color etc., perform better for scenes with corresponding pictorial depth cues. For example, depth from edge generates better depth information for scenes with high contrast or more textures. However, for scenes with lower contrast or less textures, this kind of methods will fail.

Based on motion parallax and kinetic depth effects, motion-based methods primarily use motion vectors to retrieve the depth information. Depth from motion parallax can be seen as temporal type of depth form triangular stereo vision. Depth from kinetic depth cues can estimate depth information with the phenomenon whereby the three-dimensional structure of objects viewed in projection can be perceived when the object is moving. This kind of methods can generate better depth map when effective motion exists. However, the methods fail to generate depth map when effective motion doesn't exist.

Both the pictorial-depth-cue and motion-based methods rely on that if the scene structure is suitable. From our observation, we believe that unsuitable conversion causes visual discomfort or unsatisfied 3D feeling. The 3D visual feeling of 3D videos for stereoscopic display has been widely analyzed in [1]. A number of intrinsic problems for stereoscopic display, such as cross-talk, accommodation-vergence confliction, are discussed. The intrinsic problems are defined as the causes of discomforts invoked by the display mechanism. In the other context, the recent advancements for content generation in 3D-TV systems necessitate the 3D feeling analysis of generated 3D videos. For 3D contents, researchers have discussed about the relationship between the visual quality and the compression. In [2], the authors have analyzed the effect of depth map compression on the geometric distortions for synthesized camera viewpoints. In [3], the authors have also implemented a mode decision algorithm for depth map compression in 3D-TV systems based on the visual quality on view synthesis. This paper proposes automatic depth verification based on the 3D feeling, which includes visual comfort and depth protrusion, not only for depth compression techniques but also for a variety of 3D images or videos.

In the following sections, based on the characteristics for the various types of depth generation methods, we select several videos as reference for testing the proposed metrics. Some of these sequences are more suitable for pictorial-depth-cue methods, and the others are suitable for motion-based methods. Our proposed metric will provide a quality score like PSNR. We will prove the proposed metric is highly-correlated to the subjective results.

ISSN 0097-966X/11/4203-1265-$1.00 © 2011 SID



(a) original image (b) depth from relative height [4]

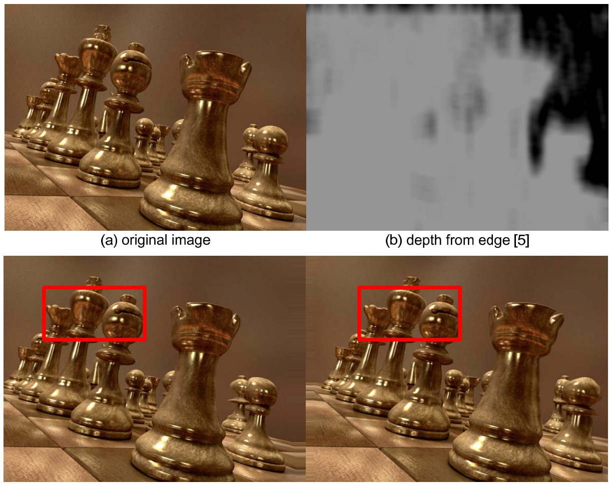
(c) synthesized left and right views base on depth (b)

Fig. 1 Observation of 2D-to-3D generated content – the same object different depth (a) original image (b) depth from relative height [4] (c) synthesized left and right views.

3. The Proposed Metrics

3.1. Relationship between Depth and Object

From our observation for various comfort and discomfort 3D sequences, we found that there is a relation between the perception of depth and formation of object. An example [4] shown in Fig. 1 illustrates the observation that pixels in the same object assigned with the different depth values may result in visual discomfort. The regions marked by the rectangles shows where visual discomfort appears. It can be observed that some pixels in the regions are within the same object but assigned with different depth values. Besides, another example [5] shown in Fig. 2 also illustrates the observation that pixels belonging to different objects are assigned with the same depth values. It may result in less depth protrusion. The regions marked by the rectangles also shows the chess and the background cannot be separated in depth, which causes less depth protrusion in 3D experience.



(a) original image (b) depth from edge [5]

(c) synthesized left and right views base on depth (b)

Fig. 2 Observation of 2D-to-3D generated content – different object the same depth (a) original image (b) depth from edge [5] (our implemented) (c) synthesized left and right views

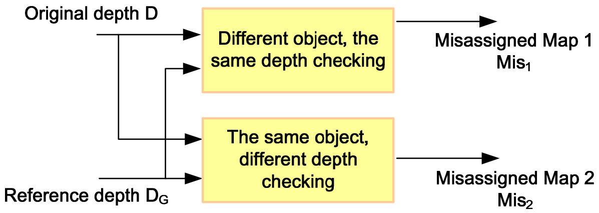

Original depth D
Different object, the same depth checking
Misassigned Map 1
Mis₁
The same object, different depth checking
Reference depth D_G
Misassigned Map 2
Mis₂

Fig. 3 The proposed assessment framework

In the following sections, we propose a full reference objective assessment framework, which tries to formulate the depth and object consistency.

3.2. Mathematical Model for Mis-assigned Maps

To illustrate the concepts, our proposed framework utilizes ground- truth depth map as the full reference as shown in Fig. 3.

Pixels with high depth similarity and proximity are generally to be classified into the same object. We try to adopt cross-bilateral filter [6] to formulate the two concepts into mathematical forms. This filter considers the similarity and proximity in the weighting averaging process. The cross-bilateral filter $f_{\mathrm{B,A}}$ between two images, A and B, is utilized as in (1), (2), and (3).

$$\mathrm{Proximity}(u,v,x,y)=\exp(-\frac{\left|(u-x)\right|^2}{2\sigma_x{}^2}-\frac{\left|(v-y)\right|^2}{2\sigma_y{}^2}), \tag{1}$$

$$\mathrm{Similarity}_A(u,v,x,y)=\exp(-\frac{\left|A(u,v)-A(x,y)\right|^2}{2\sigma_i^2}), \tag{2}$$

$$f_{B,A}(x,y)=\frac{\sum\limits_{(x_j,y_j)\in S}\mathrm{Proximity}(x_j,y_j,x,y)\mathrm{Similarity}_A(x_j,y_j,x,y)B\left(x,y\right)}{\sum\limits_{(x_j,y_j)\in S}\mathrm{Proximity}(x_j,y_j,x,y)\mathrm{Similarity}_A(x_j,y_j,x,y)}, \tag{3}$$

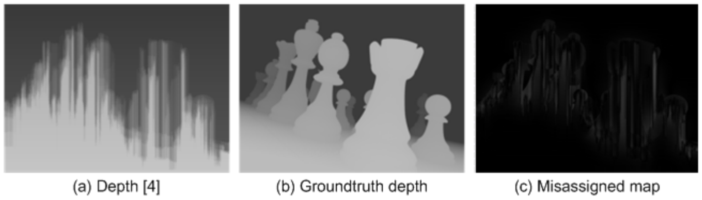
where S mean the spatial neighborhood of pixel (x, y).

The proximity function in (1) and the similarity function in (2) measure tendency of these pixels to belong to the same region with referencing the intensity of image A. The weightings for the averaging in (3) are defined by these two terms, which provide a filtered image for B by referencing to the groups in A.

3.2.1. Same Object but Different Depth (SODD)

For modeling the errors that pixels within the same object are assigned to different depth, we try to estimate the concept of object from the groundtruth depth map. We assume that the probability that pixel *(u,v)* and pixel *(x,y)* belong to the same object is proportional to the product of Proximity(*u,v,x,y*) and Similairy$_{\mathrm{G}}$(*u,v,x,y*), where G mean the referenced groundtruth depth map.

We denote D as the converted depth map by a given algorithm. By the cross-bilateral filter in (4), depth D is refined to D$_{\mathrm{Grouped}}$ on the concern of maximum likelihood. As described above, the likelihood is proportional to the product of Proximity and Similairy$_{\mathrm{G}}$. The refined depth D$_{\mathrm{Grouped}}$ will suffer from less SODD than D. Therefore, the difference map between D and D$_{\mathrm{Grouped}}$ can



(a) Depth [4] (b) Groundtruth depth (c) Misassigned map

Fig. 5 Misassigned map for the same object but different depth values (a) Depth from relative height [4] (b) Groundtruth depth (c) Mis-assigned map

indicate where SODD happens generally.

$$D_{Grouped} = f_{D,G} \tag{4}$$

For the concern of optimality for the case that D is identical to G, the mis-assigned map for SODD is constructed as (5).

$$Mis_1(x,y) = | f_{D,D}(x,y) - D_{Grouped}(x,y) |, \tag{5}$$

Figure 4 shows an example for the mis-assigned map. The regions which may cause visual discomfort by SODD are pointed out by the map.

3.2.2. Different Object but Same Depth (DOSD)

By the cross-bilateral filter in (6), depth G is destroyed to $D_{Degraded}$ by D. This filter, which is similar to the reverse operation in (4), induces some noises from D to the destroyed depth $D_{Degraded}$. $D_{Degraded}$ will suffer from more DOSD than G. Therefore, the difference map between G and $D_{Degraded}$ can indicate where DOSD happens generally.

$$D_{Destroyed} = f_{G,D} \tag{6}$$

For the concern of optimality for the case that D is identical to G, , the mis-assigned map for DOSD is constructed as (7).

$$Mis_2(x,y) = | f_{G,G}(x,y) - D_{Destroyed}(x,y) |, \tag{7}$$

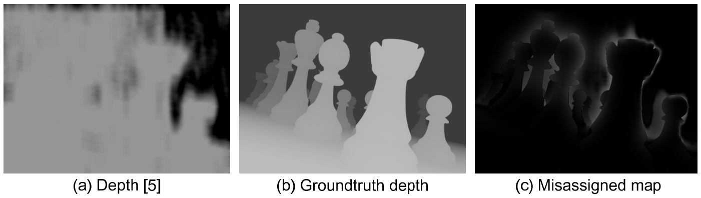
An example of the mis-assigned map is shown in Fig. 4. The regions which may cause unsatisfied 3D feeling by DOSD are also pointed out by the map.

3.3. Modified PSNR

In this section, we try to use a scalar value to represent the measured quality. First, the SODD and DOSD maps defined in section 3.2 are combined in weighted Euclidean Distance as (8) because human beings will focus their attention on pixels or elements that are unusual.

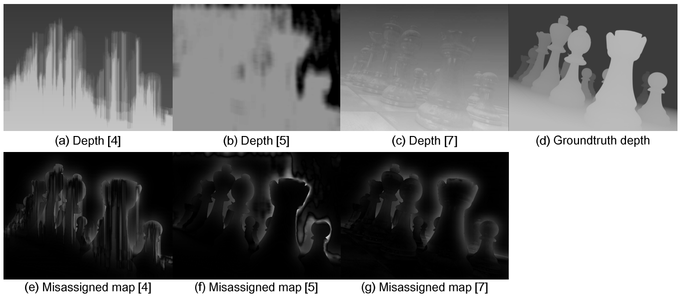
$$CombinedMis(x,y) = \sqrt{\sum_{i=1}^{2} \left(\frac{1}{1 + PeakMis - Mis_i(x,y)}\right)^2 Mis_i(x,y)^2}, \tag{8}$$

where *PeakMis* is 255 for depth maps coded in 8-bit gray-level images. The weighted term $\left(\frac{1}{1 + PeakMis - Mis_i(x,y)}\right)$ enhances the importance of unusual regions. Fig. 6 shows the combination of the proposed weighting-combined mis-assigned maps. A modified PSNR based on *CombinedMis* is defined as in (9).



(a) Depth [5] (b) Groundtruth depth (c) Misassigned map

Fig. 4 Mis-assigned map for different object but the same depth value (a) Depth from edge (our implemented) [5] (b) Groundtruth depth (c) Mis-assigned map



(a) Depth [4] (b) Depth [5] (c) Depth [7] (d) Groundtruth depth

(e) Misassigned map [4] (f) Misassigned map [5] (g) Misassigned map [7]

Fig. 6 Combined mis-assigned maps (a) Depth from relative height [4] (b) Depth from edge [5] (c) Depth from color and scene [7] (d) Groundtruth depth (e) Mis-assigned map for depth from relative height [4] (f) Mis-assigned map for depth from edge [5] (g) Mis-assigned map for depth from color and scene [7]

$$ModifiedPSNR = 10\log\left(\frac{PeakMis^2}{Average(CombinedMis(x,y)^2)}\right). \tag{9}$$

4. Experiment Results

The objective measurements by our proposed metrics are shown in Table I. For verifying our proposed method, the subjective evaluation was performed by 6 people with normal or correct-to-normal visual acuity and stereo acuity. The participants watched the stereoscopic video in a random order and were asked to rate each video according to two factors, depth protrusion and visual comfort. Sample frames of the tested sequences from [8] are shown in Fig. 7. The corresponding depth for the sequences is exploited as the groundtruth depth for our proposed metrics. The participants were also asked to give overall scores for 3D feeling based on the two factors above. The quality for depth protrusion was accessed using a five-segment scale as shown in Fig. 8(a), and that for visual comfort is shown in Fig. 8(b). The overall scores acquired by experiments for the three evaluation sequences are shown in Fig. 9. Observing Table I and Fig. 9, our proposed metrics do provide a good measurement which is highly-correlated to subjective results.

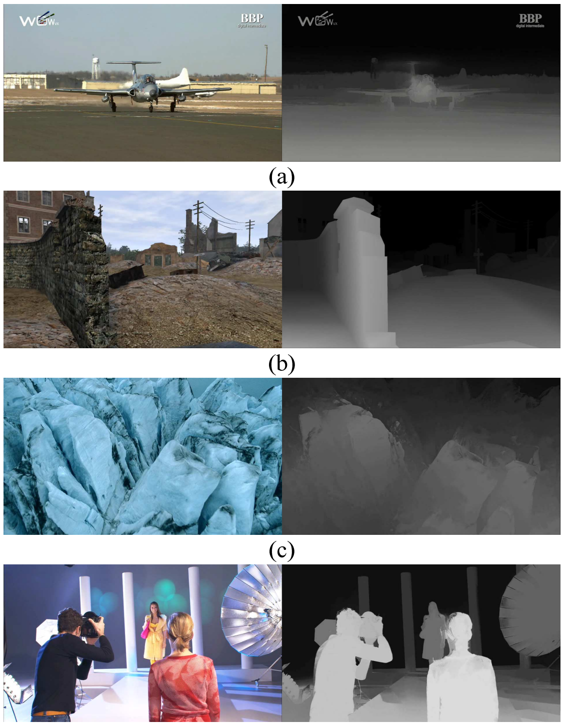

(a)
(b)
(c)
(d)

Fig. 7 Test sequences from [8]

(a) Air (b) Cod (c) Arctic (d) Fashion
Left: original frame Right: Reference depth

5. Conclusion

Compared with previous works, our proposed work utilizes the cross-bilateral filter to effectively measure the relation between the perception of depth and the formation of object. In our simulation and experiment results, we successfully detect where visual discomfort and unsatisfied 3D feeling may happen as well as human visual system does. Our work can provide an objective quality measurement for 3D images/videos to accelerate the development of 3D display systems.

6. Acknowledgements

The authors thank to Prof. Su-Ling Yeh, Shin-Yi Liao, and Yong-Hao Yang for providing perceptual knowledge to help the development of metrics. They are the members of Perception & Attention Lab, department of psychology, National Taiwan University.

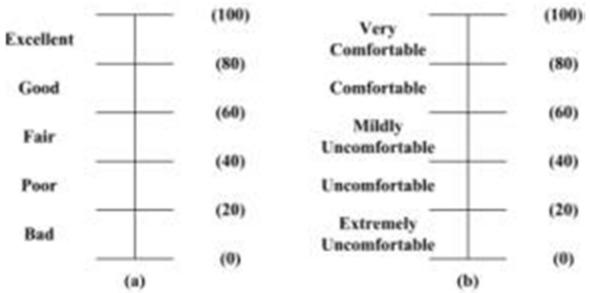


Fig. 8 - Rating scales used for assessing (a) depth protrusion and (b) visual comfort. The overall quality for depth quality was accessed using a five-segment scale and that for visual comfort.

TABLE I THE PROPOSED METRICS FOR THE TEST SEQUENCE

Sequence/ Depth Generation Method	Air (dB)	Cod (dB)	Arctic (dB)	Fashion (dB)
Edge [5] (Our implemented)	27.83	26.93	27.09	25.95
Motion [4]	26.45	27.77	26.19	24.73
Relative Height [4]	24.79	23.27	25.07	24.01
Color and scene [7]	27.59	27.32	27.14	28.11

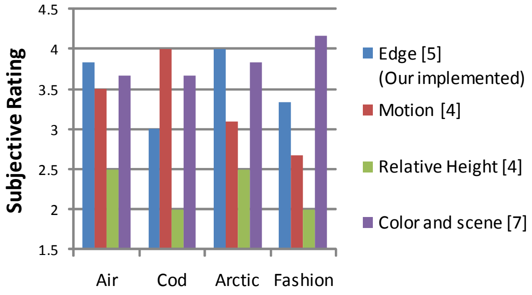


Fig. 9 Experiment Results (a) Air (b) Cod (c) Arctic (d) Fashion

7. References

[1] M. T. M. Lambooij, W. A. Ijsselsteijn, and I. Heynderickx, "Visual discomfort in stereoscopic displays: a review," in Stereoscopic Displays and Virtual Reality Systems XIV, **6490** of Proceedings of SPIE, pp. 1–13, San Jose, Calif, USA, (2007).

[2] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Mueller, P.H.N. de With, T. Wiegand, "The Effects of Multiview Depth Video Compression on Multiview Rendering", Signal Processing: Image Communication, **24**, pp. 73-88, (2009).

[3] D. De Silva, W.A.C. Fernando, H. Kodikara Arachchi, "A New Mode Selection Technique for Coding Depth Maps of 3D Video", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), (2010).

[4] Chao-Chung Cheng, Chung-Te Li, Yi-Min Tsai, Liang-Gee Chen, "A Quality-Scalable Depth-Aware Video Processing System," International Symposium, Seminar, and Exhibition of Society For Information Display(SID), (2009).

[5] W.J.Tam, A.Soung Yee, J.Ferreira, S.Tariq and F.Speranza, "Stereoscopic image rendering based on depth maps created from blur and edge information," Proceedings. of SPIE: Stereoscopic Displays and Applications XII, **5664**, pp.104-115, (2005).

[6] Elmar Eisemann and Frédo Durand, "Flash photography enhancement via intrinsic relighting," ACM Transactions on Graphics, **23**, no. 3, pp.673-678, (2004).

[7] Chao-Chung Cheng, Chung-Te Li, and Liang-Gee Chen, "An Ultra-Low-Cost 2D-to-3D Conversion System," International Symposium, Seminar, and Exhibition of Society For Information Display(SID), (2010).

[8] http://www.business-sites.philips.com/3dsolutions/