STUDY OF SUBJECT AGREEMENT ON STEREOSCOPIC VIDEO QUALITY

Ming-Jun Chen[1], Do-Kyoung Kwon[2], Alan C. Bovik[1]

[1]Laboratory for Image and Video Engineering (LIVE), Department of Electrical & Computer Engineering, The University of Texas at Austin, USA.
[2]Systems and Applications R&D Center, Texas Instruments
12500 TI Blvd., Dallas, TX 75243

ABSTRACT

We describe a study that aims towards enhancing our understanding of the perception of H.264/AVC compressed stereoscopic 3D videos, in particular spatial video quality, depth quality, visual comfort and overall 3D video quality. The results of this study indicate that the human subjects have diverse opinions on depth quality scores but a high agreement on spatial video quality. Their agreement on overall 3D video quality is intermediate relative to that on spatial video quality and depth quality. Based on our analysis, we propose to use separate quality assessment models: spatial video quality models and depth quality models.

***Index Terms*—** 3D video quality, depth quality, comfort visual, 3D video database, psychometrics

1. INTRODUCTION

As stereoscopic display technologies have advanced, stereoscopic 3D content has become quite popular. Research on methods for automatic quality assessment of stereoscopic 3D content is a hot topic and the design of effective quality assessment indices is highly anticipated. Towards this end, it is important to understand and model the human perception of *distortions* in 3D content.

Human studies on the various aspects of the perception of stereoscopic 3D content have been conducted for decades. For example, in [1], it was claimed that the binocular sense of the quality of asymmetric MPEG-2 distorted images is approximately the average of the quality of the two views, but that the perception of asymmetric blur distorted images is dominated by the higher quality view. In [2], it was claimed that the subjective quality score of a stereo sequence is approximately the average of both views when MPEG-2 distortion is applied. The authors of [3] [4] support previous findings on JPEG compression distortions, claiming that JPEG encoding has no effect on perceived depth. However, the authors of [5] claim that perceived depth is correlated with stereo content quality. This disagreement raises a basic question: "Is there a general agreement on the quality of stereoscopic 3D content across subjects?"

Table 1 The QP values for the left view and right views of the stereo 3D video

Left view QP	Right view QP
25	Pristine
30	Pristine
35	Pristine
25	25
30	25
35	25
30	30
35	30
35	35

In addition, Seuntiëns [6] proposed to evaluate the 3D viewing experience by combining three different quality assessment models: image quality, depth quality and visual comfort. In this paper, we consider whether and in what manner depth quality changes with content quality. Second, we consider whether a single "quality of experience"(QoE) model can capture the overall 3D viewing experience, or whether separate models are needed to describe different aspects of the 3D QoE. Specifically, we would like to know which subjective quality scores should be incorporated into a single stereo 3D quality database or whether separate databases are needed to study different aspects of 3D QoE. Further, we discuss the way these subjective quality scores interact.

In the following sections, we report a study on the human perception of spatial video quality (SVQ), depth quality (DQ), visual comfort (VC) and overall 3D video quality (3DVQ) using a matched-pairs experimental design.

2. METHOD

2.1. Stimuli

Six uncompressed natural scene videos, including indoor and outdoor scenes, were chosen as source videos. Two of them (soccer, puppy) are from the ETRI in Korea and the

978-1-4673-1830-3/12/$31.00 ©2012 IEEE

other four are from the EPFL stereo video database [7]. All videos were down-sampled to 720 x 480 resolution. Two of these videos are fifteen seconds long, while the rest are ten seconds long. All of the sequences have a frame rate of 25 frames per second.

H.264/AVC compression was chosen as the distortion method and an asymmetric coding scenario was included. Each pristine sequence was used to create 9 distorted test sequences compressed with different quantization parameter (QP) values. The specific settings for the nine distorted videos associated with each original video are shown in Table 1.

2.2. Display

An nVidia active 3D kit plus an Alienware OptX AW2310 full HD 3D monitor were used to display the 3D videos. The viewing distance from subjects to screen was fixed at 23 inches which is 3 times the screen height.

2.3 Study design

We adopted a single stimulus continuous quality scale (SSCQS) protocol to obtain subjective quality ratings for all of the video sequences in the database. A training session was given to each subject at the beginning of the study to familiarize them with the graphical user interface (GUI). The subjects were pre-screened to ensure normal stereovision by asking subjects to identify 2D and 3D content in the training section. In addition, a pristine video and a "most distorted" video were shown in the training session to help observers normalize their ratings. The training content was different from the videos used in the study and the content was impaired by the same type of distortion. Repeated viewing of the same 3D video was allowed, since we found that subjects sometimes needed time to accommodate their eye convergence to a new 3D video.

The goal of this work is to understand subjects' ratings of 'spatial video quality', 'depth quality', 'visual comfort', and 'overall 3D video quality'. However, in experiments preliminary to the study we found that it was difficult for subjects to rate these quality scores independently. Further, when being asked to give an overall 3D quality score for each stimulus, subjects tended to have trouble assigning relative 'weights' to SVQ, DQ, and VC. Hence, a matched-pairs experimental design was used to conduct the study.

In the matched-pair study, the study is repeated using two groups of subjects to obtain matched measurements of subjective scores. In the first study, the subjects in group A were requested to give subjective scores on SVQ, DQ, and VC. In assigning SVQ, the subjects were requested to assign quality scores only based on the content quality they viewed without considering the quality of their 3D viewing experiences. In addition, the subjects were asked to assign depth quality scores based only on the amount of 3D depth they viewed when viewing stereo 3D videos. The subjects were also asked to give a visual comfort score based on how comfortable they felt when viewing stereo 3D videos. In the second study, the subjects in group B were requested to give an overall 3D quality score when viewing stereo 3D videos. Again, the task of rating videos was explained carefully in the training session prior to each subjects' participation. Instructions were given to observers so that the scoring is based on overall 3D viewing experience.

In both study groups, 11 video sequences (a 3D pristine video, a 2D pristine video (right view), and nine distorted videos) were shown to the subjects for each pristine video. The 3D reference video was hidden to enable the calculation of DMOS scores of perceived spatial video quality and overall 3D video quality.

Subjects having similar backgrounds were recruited for the two groups. In group A, thirteen subjects (twelve males and one female) were recruited with ages ranging from 24 to 45. In group B, fourteen subjects (eleven males and three females) were recruited and their ages ranged from 24 to 50.

2.4 Obtaining subjective scores

Differential mean opinion scores (DMOS) were calculated by subtracting the ratings of each 3D reference video from each associated rating. Those scores were then normalized to Z-scores. Outliers were removed by tossing out any ratings falling outside two standard deviations from the center of a Gaussian fit to the ratings' SROCC against mean DMOS. Finally, the DMOS score of each video was computed as the mean of the rescaled Z-scores from the remaining subjects following subject rejection. After the subject rejection process, only one subject was rejected in group A. No outlier was found in group B.

3. DATA ANALYSIS AND DISCUSSION

3.1 Within Quality Assessment Metrics

Following Seuntien et al. [3], we calculated the standard deviation of the normalized ratings (Z-scores scaled to 0~100) assigned to each video. The average of these standard deviation values was then used to represent the degree of variation of the ratings. Table 2 shows that the ratings assigned to the perceived spatial video quality exhibit the least variation, although the standard deviation does not reveal the degree of agreement between the ratings. Since we are more interested in the agreement (in the sense of relative rankings, not absolute values) between ratings given by different subjects, we used the correlation between the ratings given by different subjects to discover whether their ratings were similar across the four kinds of 'qualities.' We first calculated the correlation values between the mean scores and the ratings given by each subject. The average of these correlation values reflects the degree of agreement of ratings among the subjects.

Table 3 shows the Spearman Ranked Order Correlation Coefficients (SROCC) and the Pearson Correlation

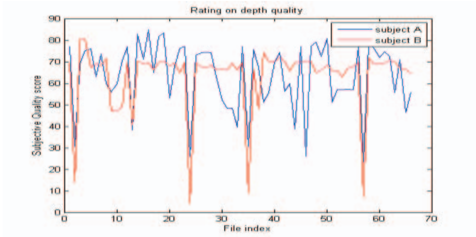

Fig. 1 Ratings of depth quality from two distinct subjects.

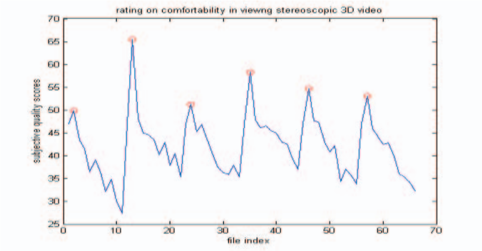

Fig. 2 Mean rating of comfort when viewing stereoscopic 3D video. The red circle represents 2D videos.

Coefficients. The ratings of SVQ show the highest agreement while the ratings of DQ show the least.

To further analyze the low agreement ratings for DQ, the data shows that some subjects assigned a lower depth quality score when the video had lower spatial video quality, while others subjects thought that compression distortion did not affect perceived depth quality. Fig. 1 is an example that shows the rating of two subjects in our study. Subject 1 assigned a variety of depth quality scores while subject 2 assigned very similar depth quality scores. Across multiple subjects, there were diverse options in interpreting depth quality. Discovering why different people have different opinions of depth quality is worthy of further exploration.

The degree of agreement of ratings on overall 3D video quality is lower than on spatial video quality and higher than on depth quality. This observation may provide insight on how to build a 3D video quality database.

The ratings of visual comfort assigned when viewing distorted stereoscopic 3D videos show a middle degree of agreement. Given that the underlying 3D geometric setting of the distorted videos is unaltered and carefully dealt with to ensure no accommodation-vergence conflict and crosstalk caused by the viewing setting, any discomfort in viewing a stereoscopic video resulted either from the intrinsic geometry of the videos or the compression distortion. Although subjects did not closely agree on visual comfort, our data show that they were more comfortable when viewing the hidden 2D pristine video. As shown in Fig. 2 the subjective scores assigned when viewing 2D video were the highest comfort scores.

Our possible explanation for the phenomena we have observed is that human beings are more familiar with distortions in 2D videos than in 3D videos. Television was invented in the late 1930s and we have been living with distorted 2D videos for a long time, whereas, for most people, stereoscopic 3D video viewing is still a new experience. Viewing stereoscopic 3D videos is a much more complex task than daily stereo vision where the eyes verge and focus at the same time. However, when viewing stereoscopic 3D video, the two eyes only change vergence while the focused point is fixed on the screen. So, most human subjects may be insufficiently experienced in viewing stereoscopic 3D video to reliably judge perceived depth quality. This may partly explain why the subjects had more diverse opinions on perceived depth quality and why they felt more comfortable viewing 2D videos. Lastly, humans exhibit a wide range of stereoacuity and stereosense [8], ranging from complete deficiency to better than normal. This ability would naturally affect a subject's impressions of both 3D distortions and comfort.

3.2 Correlating Quality Assessment Metrics

In this section, the interactions between the subjective quality metrics are discussed. Table 4 shows SROCC scores between these subjective quality metrics. First, the subjective quality metric has high correlation with visual comfort and overall 3D quality. The results indicate that visual discomfort mainly results from coding artifacts since other variables are controlled in this study, and the overall 3D quality is more correlated to spatial quality than to depth quality, as mentioned in previous work [6]. Second, for depth quality, this subjective measurement doesn't have a high correlation with spatial quality and visual comfort, but it is correlated with overall 3D quality. Finally, both visual comfort and overall 3D quality are most correlated with spatial quality.

3.3 Discussion

Seuntiëns [6] proposed that the 3D visual experience can be predicted by combining spatial quality and depth quality. From our results, since visual comfort is highly correlated with spatial quality, overall 3D quality should be able to be predicted only from spatial quality and depth quality. A linear regression was performed to verify this model with our data. The predictive model is shown in the following:

$$\overline{Y} = a \cdot SVQ + b \cdot DQ + c \cdot VC + d,$$

where $\overline{Y}$ is the predicted overall 3D quality and d is a constant. Following linear regression, the SROCC between $\overline{Y}$ and actual overall 3D quality is 0.905, which is higher than using only spatial quality to predict overall 3D quality. The regression coefficients have value a = 0.65, b = 0.32, c = 0.35 and d = -17. However, a simpler model using only *SVQ* and *DQ*:

$$\overline{Y}' = a \cdot SVQ + b \cdot DQ + d$$

can achieve the same performance: the SROCC between

Table 2 Mean of standard deviations of ratings.

	Average std of ratings
Spatial Quality	8.56
Depth Quality	10.26
Visual Comfort	8.327
Overall 3D Quality	12.98

Table 3 Mean of correlations.

	Mean SROCC	Mean Correlation
Spatial Quality	**0.806**	0.829
Depth Quality	**0.549**	0.549
Visual Comfort	0.627	0.657
Overall 3D Quality	0.644	0.706

Table 4 SROCC between subjective quality metrics.

	SVQ	DQ	VC	3DVQ
SVQ	1	0.522	0.891	0.844
DQ	0.521	1	0.429	0.686
VC	0.891	0.429	1	0.765
3DVQ	0.844	0.686	0.765	1

Table 5 SROCC of PSNR and MS-SSIM against spatial quality and overall 3D quality

	PSNR	MS-SSIM
Spatial Quality	0.790	0.820
Overall 3D Quality	0.769	0.675

$\overline{Y}'$ and overall 3D quality is 0.90 and the regression coefficients are a = 0.80, b = 0.64 and d = -6.8. Thus, we verify that the 3D viewing experience can be predicted using a single linear model from spatial quality and depth quality.

However, as we can see, overall 3D quality has a significantly lower agreement (0.644) among subjects compared to the agreement (0.86) of spatial quality. This finding suggests that we should use two independent quality assessments models: a spatial quality metric and depth quality metric, to evaluate the quality of 3D content for the purpose of providing more reliable results. For applications such as 3D content encoding and 3D content broadcasting, the geometry used in creating the content won't be altered during the encoding or transmission. Only the distortion caused either by insufficient bit-rate or packet lost will lower the content quality. Hence, using the subjective spatial quality scores to evaluate quality assessment models will provide us more reliable results across subjects.

Table 5 shows the SROCC number of two quality assessment metrics, PSNR and MS-SSIM, evaluated by two different subjective quality scores: spatial quality and overall 3D quality. The quality scores of the 3D content is simply the average of the predicted quality scores from both views. In Table 5, MS-SSIM has the better performance if the QA metrics are evaluated against spatial quality, but it performs worse if 3D quality is used for verification. However, previous work pointed out that the MS-SSIM significantly outperforms PSNR when evaluating 2D content quality.

4. CONCLUSION

Our study shows that humans tend to agree on spatial video quality, but have more diverse opinions on depth quality. The agreement on overall 3D quality scores is intermediate compared to video quality and depth quality. Although overall 3D quality can be predicted by combining spatial quality and depth quality, it provides significantly less reliable results across subjects as compared to spatial quality. Hence, instead of using one overall 3D quality model, we propose to use two independent quality models: a spatial quality model and a depth quality model to evaluate the quality of 3D content. This approach can provide more reliable QA assessment results for applications that relate to spatial quality in 3D content. For depth quality, this work didn't discuss the situation where the models of the 3D content are altered. Going forward, more human studies are needed to deepen our knowledge of human perception of depth quality in stereo 3D content.

5. REFERENCES

[1] 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 7,* 143-153, 2001

[2] L. B. Stelmach and W. J. Tam, "Stereoscopic image coding: effect of disparity image quality in left- and right-eye views," *Signal Processing: Image Communication 14,* 111-117, 1998.

[3] P. Seuntiëns, L. Meesters, and W. Ijsselsteijn, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation," *ACM Trans. Appl. Percept.,* vol. 3, no. 2, pp. 95–109, 2006.

[4] L. Meesters, W. Ijsselsteijn, and P. Seuntiëns, "A survey of perceptual evaluation and requirements of three dimensional TV," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 14, no. 3, pp. 381-391, Mar. 2004.

[5] W. J. Tam, L. B. Stelmach, and P. J. Corriveau, "Psychovisual aspects of viewing stereoscopic video sequences," *Stereoscopic Displays and Virtual Reality Systems V,* 3295, pp.226-235, 1998.

[6] Pieter J.H. Seuntiëns, "Experience of 3D TV," *PHD thesis*, 2006.

[7] L. Goldmann, F. De Simone, T. Ebrahimi: "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," *Electronic Imaging (EI), 3D Image Processing (3DIP) and Applications, San Jose, USA, 2010.*

[8] C. M. Zaroff, M. Knutelska, T. E. Frumkes. "Variation in stereoacuity: normative description, fixation disparity, and the roles of ageing and gender," *Invest. Ophthalmo.l Vis. Sci.,* vol 44, pp. 891-900, 2003.