A Subjective Comparison of Depth Image Based Rendering and Frame Compatible Stereo for Low Bit Rate 3D Video Coding

Peshala Pahalawatta and Kevin Stec
Dynamic Digital Depth, Inc., Los Angeles, CA
E-mail: ppaha@ddd.com, kstec@ddd.com

***Abstract*— Frame compatible stereo video delivery has become a de-facto standard because it enables the delivery of stereoscopic information over legacy devices that can currently only decode a 2D signal. At the cost of reducing spatial resolution of the images, frame compatible delivery also reduces the bandwidth requirements for signaling stereoscopic 3D video. The new generations of playback devices are less constrained than legacy devices in that they are increasingly becoming capable of decoding multiple video streams in parallel. Bandwidth, however, remains an issue especially in mobile wireless and real-time streaming environments. This paper explores the use of texture and depth data to render 3D views, and compares the bandwidth requirements of the depth based rendering method to frame compatible stereo. Some interesting subjective observations that affect the comparison are discussed along with the results of a formal subjective evaluation. The relative merits and drawbacks of each method are detailed both in terms of compression efficiency and overall quality of experience.**

I. INTRODUCTION

Since the advent of 3D capable displays and an increasing number of stereoscopic 3D movies, the delivery of 3D content to home viewers has been of interest to content providers. Most current 3D delivery services such as satellite and cable broadcasters, internet on-demand content providers, etc., have adopted frame compatible 3D video delivery as a de-facto standard [1]. Frame compatible schemes, which combine the left and right eye view images into one image prior to encoding, result in reducing the resolution of the images for each eye by a factor of two but do enable encoding, transmission and decoding using legacy equipment and software. For example, a video decoder capable of decoding a 1920x1080 resolution video stream at 24 frames/sec can be used to decode a side-by-side formatted frame compatible stereoscopic image sequence of the same frame rate where each eye takes up 960x1080 resolution. At the cost of reduced resolution, the frame compatible schemes also reduce the bandwidth required to transmit 3D images compared to transmitting both views at full resolution.

Increasingly, however, video decoding and playback devices are gaining processing power. As a result, it can be expected that the next generation of devices will be capable of decoding more than one video stream at a time. Furthermore, with on-demand content, the importance of backward compatibility of bitstreams diminishes since the server can choose the required transmission format depending on the available equipment. Therefore, a number of alternative schemes have been, or are in the process of being developed for 3D video coding and delivery. They include MVC - the multiview extension to the AVC standard [2], the 2D (i.e., monoscopic) plus depth coding scheme [3], layered depth video [4], and multiview plus depth coding schemes [5][6].

The limitations on bandwidth, however, will still be an important consideration due to the need for real-time applications such as real-time video streaming, cloud gaming, etc., as well as the need to provide content over limited bandwidth networks such as mobile wireless networks.

Taking the above into consideration, we investigate the use of 2D (i.e., monoscopic video) plus depth coding [2] to enable the transmission of 3D video over bandwidth constrained networks. As detailed in [3], 2D plus depth coding has a number of advantages as well as some disadvantages compared to stereo video coding when used for 3D video. 2D plus depth coding requires less bandwidth than frame compatible schemes and less decoder complexity than the multi-view plus depth schemes that are currently being considered such as in the 3DV project in MPEG [5]. Therefore, it is likely that the 2D plus depth scheme is a viable alternative, especially in the case of limited bandwidth applications where the playback devices are also limited in computational power. To our knowledge, however, there is little or no existing data that quantifies the tradeoff in performance between frame compatible schemes and 2D plus depth schemes for 3D video coding in terms of user experience. The goal of this paper is to subjectively evaluate the performance of the two schemes at low bit rates, and to help better understand the advantages and disadvantages of each scheme.

The rest of the paper is organized as follows: in Sec. II, we briefly describe each coding scheme, and discuss the known advantages and disadvantages of each scheme. Readers who are intimately familiar with the two schemes may skip to Sec. III where we describe the subjective testing methodology used for the formal evaluation of the two schemes. In Sec. IV, we show the results of the subjective evaluation and discuss some of the interesting findings. We conclude in Sec. 0 with some ideas for future work.

II. 3D VIDEO CODING SCHEMES

This section provides a brief overview of the two coding schemes that were evaluated in this paper. We include some of the known advantages and disadvantages of each scheme.

A. Frame Compatible Stereo Coding

An excellent overview of the frame compatible stereo coding schemes is provided in [1], and therefore, this paper will not elaborate on the different frame compatible stereoscopic image formats. Essentially, such formats tend to reduce the resolution of the original left and right eye views by half. Typically, the resolution reduction is performed by subsampling the left and right eye images either vertically as in the Top-and-Bottom (TaB) format, horizontally as in the Side-by-Side (SbS) format, or diagonally as in the quincunx sampling format. Another frame compatible method is the tile format [7], which further reduces the overall pixel resolution (1.5:1 in each dimension in the case of [7]) but provides some advantages such as potential compatibility with 2D display devices (depending on the capabilities of the playback device) and lower loss of resolution along a single dimension. For the purposes of this paper, we have considered the TaB and SbS formats since they are currently the most commonly used in industry for frame compatible stereo delivery.

Below are some advantages of the frame compatible schemes:

- Compatible with legacy playback devices (note that compatibility in this case refers to the ability to correctly decode the signal)
- Decoder complexity equal to that of single 2D decoder
- Lower bandwidth requirements than full resolution stereo video coding
- Lookaround ability – compared to the 2D + depth scheme, stereoscopic signaling has the advantage that, assuming the original content was created as stereoscopic 3D, parts of objects that are occluded in one view will still be visible in the other view.

There are, however, a number of disadvantages also associated with frame compatible schemes. Among them are:

- Not directly compatible with 2D displays – Additional processing must be performed at the playback device in order to view a 2D image using the frame compatible signal.
- Higher bandwidth requirements than 2D signal – Although in terms of the pixel resolution of the coded image, the frame compatible signal is equivalent to a 2D signal, the actual bandwidth requirements to achieve the same fidelity of the image tend to be somewhat higher. This is caused to a large extent by the loss of spatial correlation due to image subsampling. The increase in bandwidth can be controlled to some extent by appropriate low-pass filtering prior to subsampling but for detailed images can still be in the order of 20-30% of coding one eye at full resolution with the same fidelity. To a smaller extent the increase in bandwidth can also be caused by the loss of correlation across the view boundaries in the image. Also, frame compatible schemes that are backward compatible with legacy devices cannot exploit inter-view correlation that exists between the two views.
- Less control of display adaptation – Stereoscopic signaling in general reduces the ability to allow the user to adapt the content to a particular display size, or viewing distance. This can result in viewer discomfort when viewing stereoscopic images that have been tuned to substantially different display parameters. Providing the ability for display adaptation requires additional complexity at the decoder, which includes the use of automatic depth-from-stereo algorithms that are prone to error.
- Lower correlation between views – Since each view is essentially coded independently, frame compatible schemes can result in uncorrelated coding artifacts between the two views. Interesting effects of this phenomenon were noticed during this study where uncorrelated compression noise resulted in visible and annoying depth artifacts in the frame compatible stereo images when viewed in 3D.

B. Depth Image Based 3D Video

An overview of the depth image based rendering scheme for 3D video coding is provided in [8]. The scheme relies on the transmission of a monoscopic "source" image, and a corresponding depth image that can be used to generate additional viewpoints of the scene. As elaborated in [8], in the case of generating stereoscopic viewpoints where the camera setup is assumed to be parallel, the rendering can be efficiently performed with a one-dimensional image warping process. In that case, each pixel (u, v) in the monoscopic image will be warped to a corresponding pixel (u^*, v) in the left or right image, where u^* is found as:

$$u^* = u + \frac{\alpha t_x}{Z(u,v)} + h \,. \tag{1}$$

α is a constant scaling factor depending on the individual camera parameters, t_x is equal to half the baseline distance between the two cameras and is negative for the left eye camera and positive for the right eye camera. $Z(u, v)$ is equal to the depth at position (u, v) and h represents the horizontal sensor shift in the parallel camera setup. The value of h can be computed based on the convergence distance, Z_c, as:

$$h = \frac{-\alpha t_x}{Z_c} \,. \tag{2}$$

This value of h ensures that pixels that occur along the plane of convergence (i.e., $Z = Z_c$) will not be warped.

The advantages of the depth image based rendering techniques, especially when compared to frame compatible stereo, are:

- Compatible with 2D displays – The source image can be displayed as is on a 2D display.
- Display adaptation – The users can control both the overall scene depth and the plane of convergence to obtain a more comfortable viewing experience with no additional processing.

- High correlation between views – Since all of the views are generated using the same source image, there is high correlation among views, which reduces the visibility of compression artifacts.
- Support for autostereoscopic/multiview displays

The main disadvantage of the scheme is that the image warping process can be affected by disocclusions where portions of an object that were not visible in the monoscopic image should have become visible in the stereoscopic left or right eye view. The warping process typically operates by resampling the monoscopic view and does not have access to additional information that accounts for the disocclusion. A number of methods exist for minimizing the adverse effects of disocclusion. Among them are hole-filling techniques that interpolate or in-paint from neighboring regions, and simpler depth pre-processing techniques that filter the depth image to avoid strong discontinuities [8][9]. It has been shown that even simple depth pre-processing strategies can help reduce the more visibly annoying artifacts caused by disocclusion. Note, however, that these techniques cannot accurately reproduce the missing information, and therefore, the 2D + depth scheme cannot provide proper "lookaround" ability. Extensions to the scheme such as layered depth video schemes [4], as well as multi-view plus depth coding schemes [6] have been proposed as solutions to this problem. Due to the additional bandwidth requirements of such solutions, however, they are not considered within the scope of this paper.

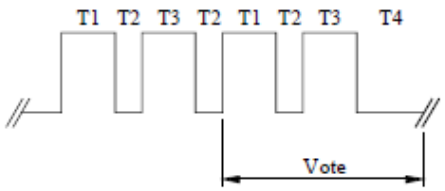
Another disadvantage of the scheme when compared to frame compatible stereo is the need for additional decoding of the depth maps, and subsequent rendering. Note, however, that the depth maps can be represented as monochrome images, and may also be reduced in resolution without significantly affecting the quality of the final rendered image. Overall, for typical content, the depth maps can be transmitted with very little bit rate overhead (less than 20%) over the corresponding monoscopic signal. The lower rate implies less complexity for entropy decoding and the lower resolution implies lower memory and processing requirements compared to decoding a typical full resolution HD image.

III. SUBJECTIVE TEST

This section details the test method and setup for the subjective viewing tests that were conducted in order to determine the performance of the two coding schemes at low bit rates.

A. Test Method

The subjective test was conducted under home viewing conditions in a controlled environment. The guidelines recommended in ITU-R BT. 500 [10] were used as much as possible within the limitations of the available equipment and resources. Rec. 500 recommends the use of the Double Stimulus Continuous Quality Scale (DSCQS) method for stereoscopic image coding tests. According to the DSCQS method, each clip is presented along with the corresponding reference clip twice prior to a voting period in which the observer notes down the scores for the pair of clips.



Phases of presentation:

T1 =	10 s	Test sequence A
T2 =	3 s	Mid-grey produced by a video level of around 200 mV
T3 =	10 s	Test sequence B
T4 =	5-11 s	Mid-grey

Fig. 1: Presentation sequence in DSCQS (excerpted from [10])

Fig. 1 shows the recommended presentation sequence according to Rec. 500 that was used in this test. The position of the reference (i.e., whether it was clip A or clip B) was randomized throughout the test. Since the test sequences were encoded at low bit rates (at or below 5Mbps), very high bit rate (20Mbps per eye) compressed stereoscopic sequences were used as references for the test. The compression of the reference was mainly necessary due to playback limitations of the player used for 3D viewing.

The observers were asked to grade each clip on a continuous grading scale which included labels designating the quality levels: "Excellent", "Good", "Fair", "Poor", and "Bad". During the training session, users were asked to evaluate each clip taking into account a combination of characteristics including the reproduction of detail, reproduction of colors, brightness and depth, as well as imperfections caused by blockiness, blurring or noise.

B. Test Setup

The test was conducted in two different locations with a somewhat different test setup in each location. The first location used a 40" Samsung LED 3D display (Model ES7500F) while the second location used a 50" Panasonic VT20 plasma display. Both displays used active shutter glasses for full resolution 3D viewing. The viewing distance was set to 3H (3 x picture height). The total number of observers was 28 (7 female and 21 male). All observers were of normal or corrected to normal vision and were tested for stereoscopic vision. Most of the viewers had some experience with 3D viewing.

Table 1: Sequences Used for Test

Sequence	Resolution	Source	Description
S1 (StreetFighter)	1920x1080 @24fps	Game capture	Ground truth depth maps with sharp transitions
S2 (NewsRoom [12])	1920x1080 @24fps	Stereoscopic capture.	Depth maps obtained using automatic stereo to depth algorithm. Some inaccuracies visible in depth map. High spatial detail in some areas.
S3 (Musicians [13])	1920x1080 @25fps	Multiview capture	Fairly accurate depth maps
S4 (Poker [13])	1920x1080 @25fps	Multiview capture	Fairly accurate depth maps

Both the stereoscopic sequences as well as the depth image based sequences were played back using the TriDef Media Player application [11]. For the monoscopic + depth sequences, the TriDef Media Player renders the corresponding left and right eye sequences using a fixed maximum disparity range and plane of convergence.

C. Sequences

Four sequences with varying characteristics were used for the test. Two of the sequences were at 24fps and two were at 25fps. They were all of 10sec duration and did not include any scene cuts, or fades. Table 1 provides a brief description of each of the sequences. All of the sequences except S1 were originally stereoscopic at 1920x1080 resolution per eye. The S1 sequence consisted of two versions, one frame compatible (TaB) stereo at 1920x1080 resolution, and the other monoscopic + depth where the monoscopic images were 1920x1080 resolution. The frame compatible stereo version of S1 was rendered using game geometry and texture data, and therefore, provided the equivalent of a stereo capture with "lookaround" ability.

It is important to consider the results of the subjective evaluation in terms of the characteristics of the content that was used for the test. Fig. 2 plots the spatial and temporal perceptual information values (SI and TI) obtained using the method recommended in ITU-T P.910 [14] for each of the test sequences. As can be seen, 3 of the sequences had relatively high spatial information and medium temporal information

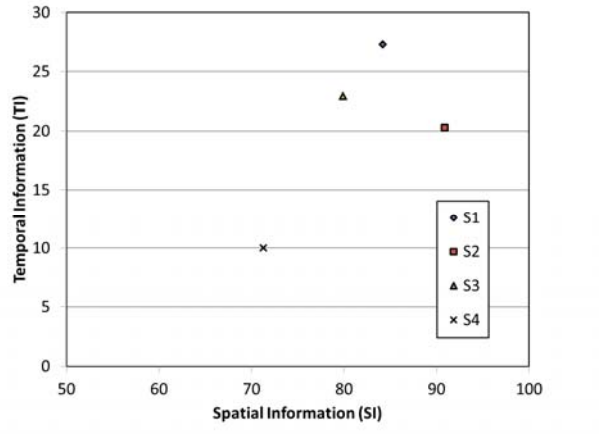


Fig. 2: Spatial and temporal perceptual information of clips

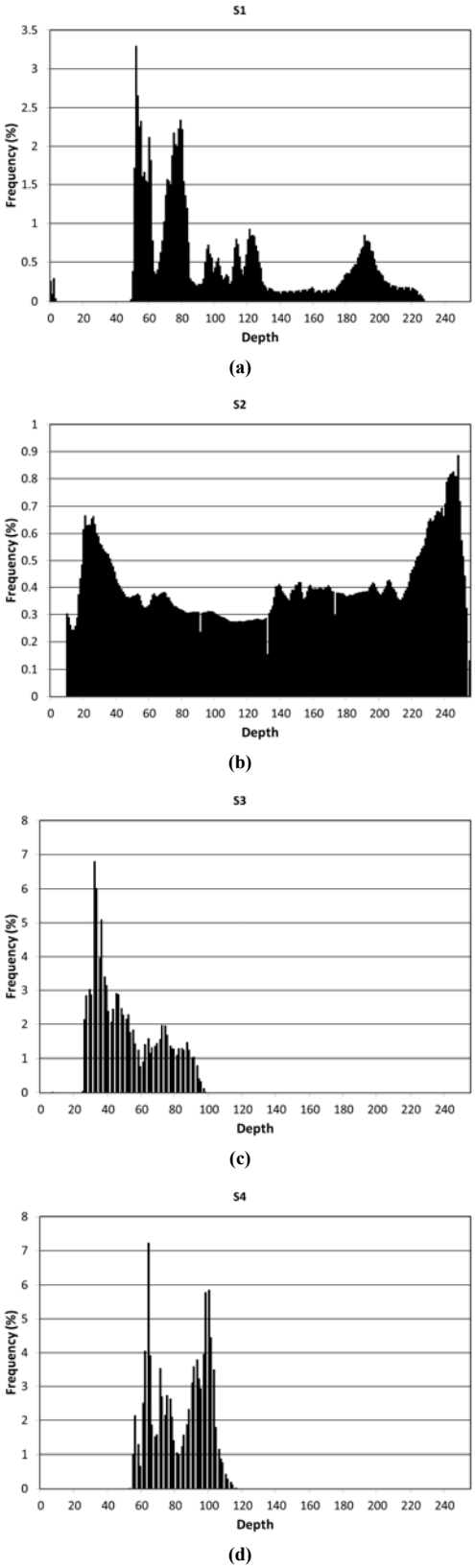


Fig. 3: Depth histograms for each sequence

while the fourth, S4, had low spatial and temporal information. S1 had the highest temporal information because it was a game capture which included a camera pan as well as moving foreground objects. S2 had high spatial information due to a striped shirt worn by the newscaster.

Fig. 3 plots the depth histograms for each of the sequences. Objects closer to the viewer received a larger value in the 8-bit depth map. As can be seen, the depth histograms for S1 and S2 show a relatively even distribution of objects in the scene in terms of the total available depth budget, while the objects in S3 and S4 were less evenly distributed. Note that S1 included the most accurate depth maps of the 4 sequences. The method used to generate depth maps for S2 resulted in some diffusion of depth to neighboring objects and explains the more continuous nature of the depth histogram in that case.

D. Test Conditions

The comparison between frame compatible stereo coding and monoscopic + depth coding was performed at 3 bit rates using each of the 4 test clips. Therefore, the total number of test conditions amounted to:

4 (clips) x 3 (rates) x 2(schemes) = 24 (tests)

Including an initial training session consisting of 4 training clips, the total time for the subjective test was approximately 30 minutes, which is reasonable to avoid viewer fatigue and discomfort due to 3D viewing.

A combination of SbS and TaB formats was used for the frame compatible stereo depending on the sequence. The bit rates tested for each sequence were chosen to provide a range of visual quality within a typical low bit rate viewing environment. The depth bit rates were chosen based on prior expert viewing to minimize rendering artifacts caused by depth compression while maintaining a maximum depth rate of 10% of the total rate. Also, based on prior viewing tests using the uncompressed source and depth, the depth maps were reduced to a quarter of the original resolution prior to encoding since that did not cause any visible artifacts in the synthesized images. The frame compatible format as well as the total bit rates and depth bit rates used for each sequence are shown in Table 2.

The encodings were performed in H.264/AVC using the x264 software [15] with two-pass rate control. All encoding parameters were kept the same for encoding the frame compatible and monoscopic sequences. For the depth images, the encoding parameters were kept the same as the others except that all perceptual optimizations were disabled since the codec is optimized for natural images and not depth images. An IBBBP coding structure and 2 sec fixed GOPs were used for all sequences. The motion estimation search range was set to 128.

IV. RESULTS

Fig. 4 thru Fig. 8 show the results of the subjective evaluation averaged over all the observers for each sequence. The results were calculated based on the difference in rating given by each observer between the reference sequence and the corresponding test sequence. Each rating, which was marked on a continuous scale, was entered as an integer in a scale of 0-100. In the graphs, we subtract the differential mean opinion score (DMOS) for each test sequence from 100 such that better visual quality is represented by a higher level along the Y-axis in each plot. The confidence intervals are calculated based on the standard deviation of the scores, assuming a student T distribution. Four outliers (two from each test location) were removed from the dataset and are not included in the results (i.e., the MOS values represent the average rating over 24 viewers).

Table 2: Encoding Formats and Rates

Sequence	FC Format	Total Rate (kbps)	Depth Rate (kbps)
S1	TaB	1500	150
		2500	250
		5000	500
S2	SbS	1000	100
		2500	250
		5000	250
S3	TaB	1000	100
		2500	250
		5000	250
S4	SbS	1000	100
		2500	100
		5000	100

The results indicate that, in general, the 2D plus depth coded content was perceived as better visual quality by the participants. The difference in performance between the two schemes increases at low bit rates for most of the sequences, which is to be expected, since the 2D plus depth scheme can be more bandwidth efficient than frame compatible coding. At the highest tested rate point, most of the sequences had similar ratings (well within the confidence intervals) for the two schemes. Sequence S3, shown in Fig. 6, is an outlier in this regard, the reason for which may be that the stereoscopic version of the sequence contained large disparities which caused visual discomfort to the viewers. At the lowest tested rate points, the improvement gained by using 2D + depth is quite significant, and as shown in Fig. 4 and Fig. 5, for the more difficult to code sequences (i.e., high SI and TI), is in the order of 20 units in the DMOS scale. This difference approximately represents a shift of an entire category label according to the 5 category labeling that was applied in the subjective test. It is also worth noting that for sequences S1 and S2, the subjective scores obtained at the low rate point for the 2D+depth scheme are statistically equivalent to those obtained at the medium rate point (i.e., 1.5x to 2x of the low rate point) for the frame compatible scheme.

The results from S4, shown in Fig. 8, are somewhat more ambiguous, most likely because of the low spatial and temporal content in the sequence. The two higher rate points were of very good visual quality in both schemes and were difficult to distinguish from each other. The lowest rate point, however, does still show a significant difference in performance. The reason for the lower score at the highest rate point for the frame compatible scheme is unclear but given the overall high scores received at the two higher rate points, may represent a plateauing effect.

One reason for the overall performance difference may be the disadvantage that frame compatible coding has in terms of the uncorrelated behavior of the left and right eye views. Compression noise in the darker regions of the scene leads to depth artifacts where a flat region may appear to have visually annoying depth discontinuities when viewed in 3D. On the other hand, the 2D + depth scheme tends to keep the noise correlated in both views, and therefore, although it is visible, the noise does not lead to depth artifacts.

Fig. 7 illustrates the inter-lab correlation between the test results from each location. It can be seen that the inter-lab correlation is acceptable with an overall correlation (R^2) of 0.83. It is likely that the difference in display types may have caused some loss in correlation between the results of the two locations.

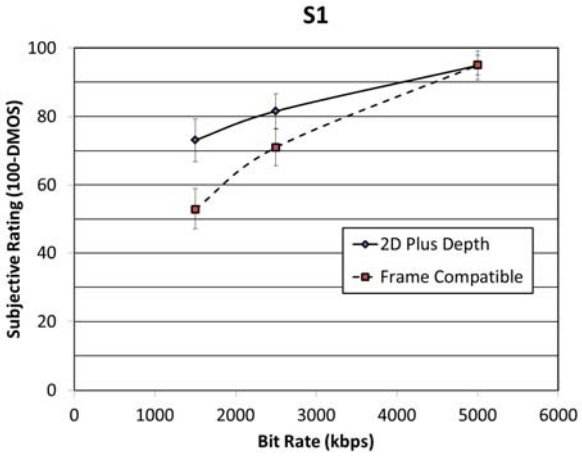


Fig. 4: Subjective comparison of FC and 2D+Depth for S1

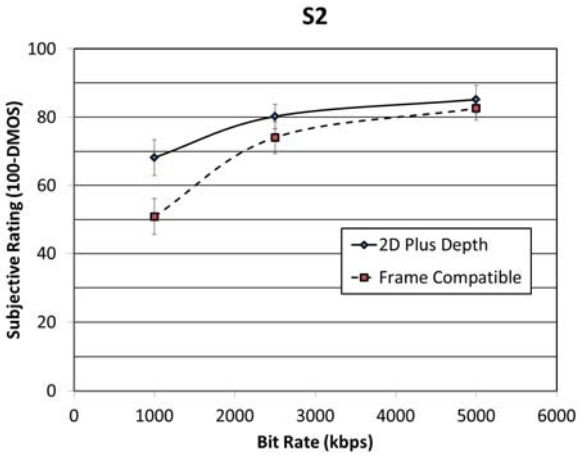


Fig. 5: Comparison of FC and 2D+Depth for S2

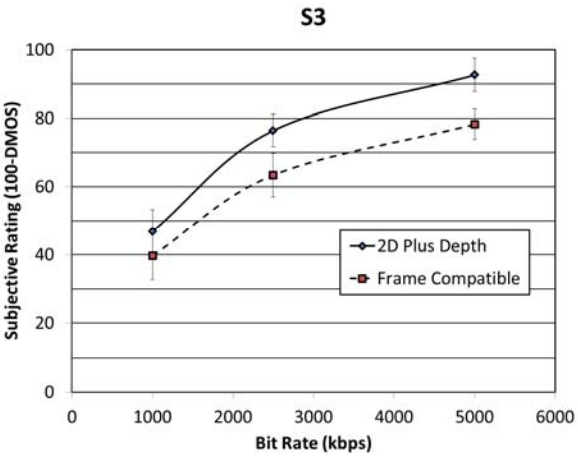


Fig. 6: Comparison of FC and 2D+Depth for S3

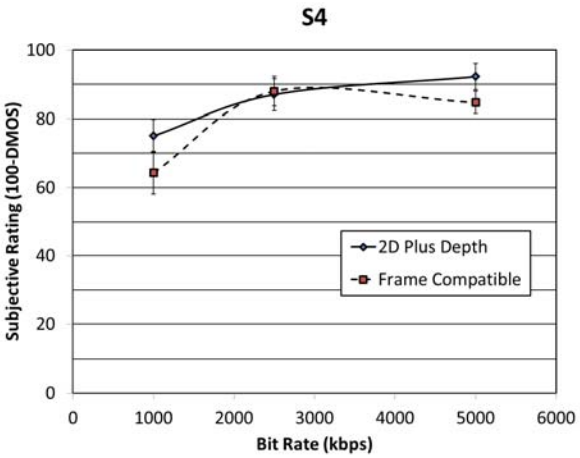


Fig. 8: Comparison of FC and 2D+Depth for S4

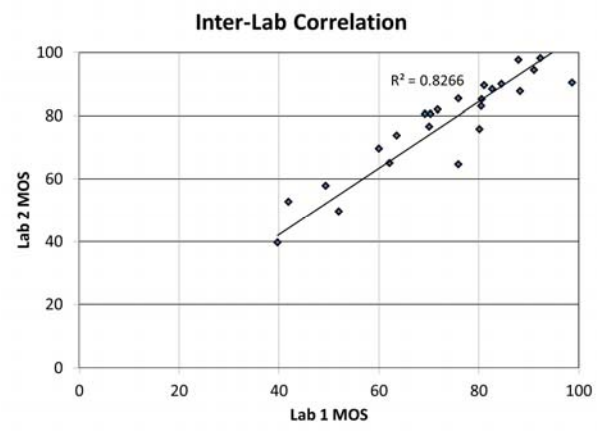


Fig. 7: Inter-lab correlation between MOS values

V. Conclusions and Future Work

This paper reports the results of a subjective evaluation conducted to determine the performance difference between frame compatible stereo coding and depth image based coding together with stereo rendering at low bit rates. The main conclusion, based on the subjective test results, is that the depth image based scheme has a statistically significant performance advantage over frame compatible stereo coding at low bit rates. The gain is visible across the variety of tested content. Overall, the results show that the 2D + depth scheme tends to degrade gracefully as the bit rate is decreased while the frame compatible scheme tends to reach an earlier breaking point at which it suffers a severe loss of visual quality. At higher bit rates, however, the performance of the two schemes tends to be statistically equivalent.

The above results are also important to consider in the context of the development of new 3D video coding standards, such as the 3DV project that is currently being conducted jointly between the MPEG and VCEG standardization groups [16]. The results indicate that monoscopic plus depth can be used as a good quality baseline to compare against new coding schemes that use stereo plus depth for multiview rendering because it performs well at low bit rates.

In the future, we plan to conduct the test on a wider range of test content and also to explore the behavior of the two schemes at higher rate points to determine if there is a crossover point where the frame compatible scheme performs significantly better than the other. We would also like to investigate the effect of the quality of the depth map, both in terms of compression artifacts and in terms of the original quality of the depth estimation, on the overall visual quality scores. It will also be interesting to include full resolution stereo coding schemes such as MVC (the multiview coding extension of the AVC standard [2]) in the test.

References

[1] A. Vetro, *Frame Compatible Formats for 3D Video Distribution,* Proc. IEEE International Conference on Image Processing (ICIP), 2010, pp. 2405-2408.

[2] ISO/IEC 14496-10|ITU-T H.264, *Advanced Video Coding for Generic Audiovisual Services,* Jan. 2012.

[3] C. Fehn, *A 3D-TV Approach Using Depth-Image-Based Rendering*, Proc. Visualization, Imaging, and Image Processing (VIIP), 2003.

[4] B. Bartczak et al, *Display-Independent 3D-TV Production and Delivery Using the Layered Depth Video Format,* IEEE Transactions on Broadcasting, vol. 57, no. 2, 2011, pp. 477-490.

[5] ISO/IEC JTC1/SC29/WG11, *Call for Proposals on 3D Video Coding Technology*, Doc. N12036, Geneva, Switzerland, March 2011.

[6] A. Smolic et al, *Intermediate View Interpolation Based on Multiview Video Plus Depth for Advanced 3D Video Systems,* Proc. IEEE International Conference on Image Processing, 2008, pp. 2448-2451.

[7] G. Ballocca, *Tile Format: A Novel Frame Compatible Approach for 3D Video Broadcasting,* IEEE International Conference on Multimedia and Expo (ICME), 2011, pp. 1-4.

[8] C. Fehn, *Depth-Image-Based Rendering (DIBR), compression, and transmission for a new approach on 3D-TV,* Proc. SPIE, vol. 5291, 2004, pp. 93-104.

[9] L. Zhang, W.J. Tam, *Stereoscopic Image Generation Based on Depth Images for 3D TV,* IEEE Transactions on Broadcasting, vol. 51, no. 2, 2005, pp. 191-199.

[10] ITU-R BT.500-11, *Methodology for the Subjective Assessment of the Quality of Television Pictures,* 2002.

[11] TriDef 3D Media Player v. 7.2.52, *http://www.tridef.com/download/TriDef-3D-latest.html.*

[12] Dolby Consumer 3D Test Sequences, Dolby Laboratories Inc., http://www.dolby.com.

[13] European FP7 Research Project Muscade (MUltimedia SCAlable 3D for Europe), *http://www.muscade.edu, grant agreement n°247010.*

[14] ITU-T P. 910, *Subjective Video Quality Assessment Methods for Multimedia Applications,* 2008.

[15] x264 build 0.122.2184, http://www.videolan.org/developers/x264.html.

[16] K. Mueller, A. Vetro, *AHG Report on 3D Video Coding,* JCT-3V Doc. JCT2-A00001, July 2012.