Integral 3D image coding by using multiview video compression technologies

Kazuhiro Hara*, Miwa Katayama*, Masahiro Kawakita*, Toshiaki Fujii*, Tomoyuki Mishina*
* NHK Science & Technology Research Laboratories, Kinuta, Setagaya-ku, Tokyo 157-8510, Japan
E-mail: {hara.k-mg, katayama.m-gm, kawakita.m-ga, mishina.t-iy }@nhk.or.jp
*Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8603, Japan
E-mail: t.fujii@nagoya-u.jp

***Abstract*— Effective compression technology is required to reduce the huge amount of information for integral three-dimensional (3D) television. For compressing an integral 3D image, we propose a compression method of converting elemental images to multiview images and of applying multiview video coding to part of the multiview images and their depth maps. In this method, the relationship between the number of the part of the multiview images and the image quality degradation of a reconstructed 3D image was studied by subjective evaluation experiment, and we confirmed the amount of information required for displaying an acceptable reconstructed 3D image. As a result, the reconstructed 3D images with acceptable image quality were obtained with about 2/9 times the amount of information for coding all the multiview images converted from the elemental images.**

I. INTRODUCTION

We are advancing research into broadcasting services for three-dimensional (3D) television and considering representation of a 3D image by using integral imaging [1-4]. Integral imaging is based on integral photography [5]. One important factor to capture and display natural 3D scene with high image quality is the number of the micro lenses on the lens array [6,7]. A micro lens of lens array is called elemental lens. And a small image captured through elemental lens is called elemental image. In integral 3D system, the number of elemental lenses is equal to the maximum resolution of reconstructed 3D image. Thus, when recording or transmitting the integral 3D image with high resolution, the data size of the elemental images apparently become huge. Therefore, more advanced technologies for video compression are required.

For the elemental images, the Moving Picture Experts Group (MPEG) has started exploration experiments on dense light field compression [8]. One topic in the dense light field compression is which video format is most suitable for achieving efficient compression performance without 3D model compression such as point cloud coding [9]. We previously developed a method to convert elemental images into multiview images and introduced multiview video coding as an effective approach for efficient compression [10]. Figure1 shows a flowchart of the compression of elemental images and multiview images captured by multiple cameras for displaying an integral 3D image. In Fig.1, when the elemental images are input, the elemental images are converted into dense multiview images. The dense multiview images are reduced part of the multiview images after estimating their depth maps, and converted to sparse multiview images. Encoder codes the sparse multiview images with their depth maps. After decoding process, the decoded sparse multiview images are used for view interpolation of the reduced number of the multiview images for reconverting the dense multiview images. At the end of process, all the dense multiview images including view synthesis images are reconverted into the elemental images. In this paper, we study the effectiveness of reducing part of the dense multiview images by assessing the visual quality of the reconstructed 3D images.

The rest of this paper is as follows. In Section II, we explain the format of sparse multiview images with the depth maps to generate the elemental images and see how much the number of images can be reduced by applying a view synthesis technique to computer graphics (CG) content. In Section III, we carry out coding experiments on the sparse multiview images generated in Section II. In Section IV, we reduce the number of the multiview images by adding the depth maps on the content captured by multiple cameras. The depth maps corresponding to the multiview images were generated by

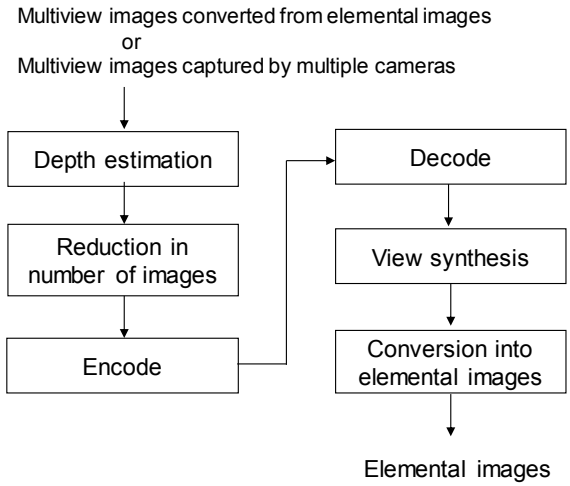

Fig. 1 Flowchart of compression of multiview images for integral 3D image.

using the latest MPEG Depth Estimation Reference Software (DERS) [11]. We coded the sparse multiview images including the depth maps by using a multiview coding called 3D-HEVC [12,13]. After rendering the synthesized images by the latest MPEG View Synthesis Reference Software (VSRS) [11,14], and converting the elemental images from the decoded sparse multiview images and the synthesized images, we subjectively evaluate the reconstructed 3D images by using the Double Stimulus Impairment Scale (DSIS) [15].

II. EFFECT OF THE REDUCED NUMBER OF DENSE MULTIVIEW IMAGES

The elemental images that display the integral 3D image can be converted into the dense multiview images. At this time, the number of pixels in the elemental image is equal to the number of the dense multiview images. Also, the resolution of the dense multiview image is equal to the number of the lenses on the display. The current integral 3D display technology will be able to make the number of lenses of the lens array reach over one hundred thousand lenses [16,17]. But, the resolution of the multiview image converted from the elemental images is still too small for the test sequence to do coding experiments. Therefore, test images from 3DCG were used instead of naturel images captured by a plenoptic camera. For rendering elemental images, Iwadate et al. developed a method for generating elemental images from 3D scene [18]. In the method, the dense multiview images captured by virtual cameras on the fixed positions in 3D scene are converted into the elemental images. In this section, we examined the effect of reducing the number of the dense multiview images to decrease the amount of information for reconstructing an integral 3D image.

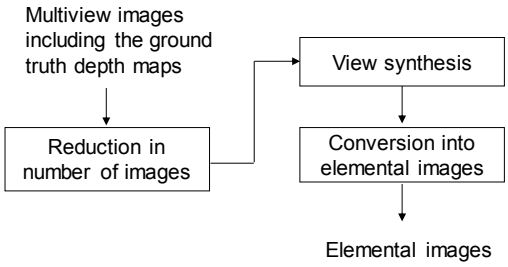

Fig. 2 Flowchart of image compression process.

A. Reduced number of dense multiview images

Figure 2 shows the image compression process in this experiment. The number of dense multiview images including the ground truth depth maps is reduced for decreasing the amount of information for reconstructing an integral 3D image and converted to sparse multiview images and their depth maps. The depth maps are also generated by using a virtual camera in 3D scene. The sparse multiview images including the depth maps are used for view synthesis, and the same number of the dense multiview images including the synthesized images are converted into the elemental images.

B. Test sequence and equipment

The test sequence is shown in Fig. 3. The specifications of the integral 3D display and the elemental images are shown in Table 1. Since the resolution of the elemental image is 21.74 × 21.74 pixels, 22 × 22 (i.e., 484) of the dense multiview images were generated. In addition, since view synthesis needs depth information for each sparse multiview image, we generated the ground truth depth maps that is expressing luminance as the distance between the virtual camera to the objects in the 3D scene. When distance is short and long, the luminance values of depth map are set to high value and low value, respectively.

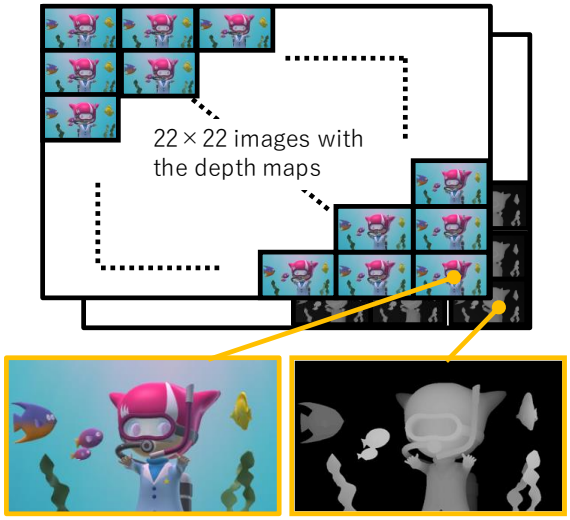

Fig. 3 Test sequence generated from CG scene.

TABLE 1
SPECIFICATIONS OF DISPLAY, LENS ARRAY, AND 3D IMAGES

Display	Size	9.6 inches
	Resolution	3840 × 2160 pixels
	Pixel pitch	55.5 um
Lens array	Focal length	2.41 mm
	Lens pitch	1.20 mm
	Arrangement	Square
	Number of lens	176 (H) × 99 (V)
3D image	Resolution of elemental image	21.74 × 21.74 pixels
	Number of elemental image	176 (H) × 99 (V)
	Size	9.6 inches
	Viewing angle	28 degrees

C. Experiment methodology

Figure 4 shows generation of anchor and evaluation images. The anchor is obtained by converting the dense multiview images (22 × 22 images) into elemental images. The evaluation images are converted from the sparse multiview images and view synthesis images, that are rendered by the sparse multiview images and the depth maps, into elemental images. And the sparse multiview images for 8 × 8, 4 × 4, and

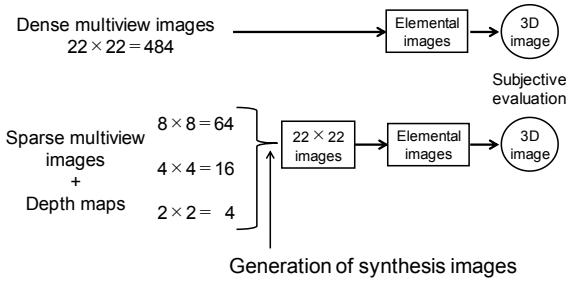

Fig. 4 Generation of anchor and evaluation images for subjective evaluation experiment.

2 × 2 images were selected at equal intervals from the dense multiview images of the 22 × 22 images. In the subjective evaluation experiment, the DSIS method Variant II specified in ITU-R Recommendation BT.500[15] was used. Figure 5 shows the experiment setting and definition of each opinion score in DSIS method.

A reconstracted 3D image was displayed and evaluated on the integral display the specifications of which are in Table 1. In the subjective evaluation experiment, an evaluator sat approximately 2.1 m from the integral display by reason of the lens pitch on the lens array put in flont of the display panel and the display size. Eight experts in video technology participated in the experiment.

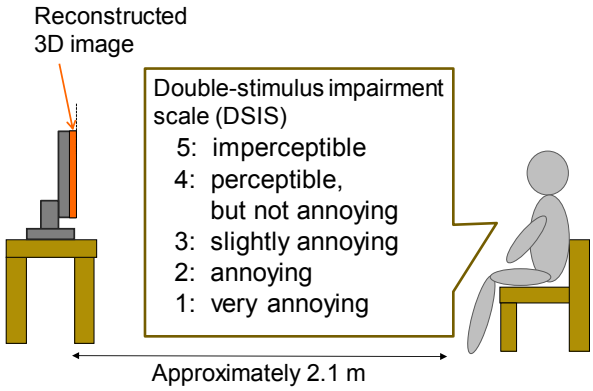

Fig. 5 Experiment setting and definition of each opinion score in DSIS method.

D. Results

The relationship between the amount of data for a reconstructed 3D image and the mean opinion score (MOS) is shown in Fig.6. The data size along the x axis is normalized by the data size of the multiview images of 22 × 22 images. In Fig. 6, the amount of data for depth map is estimated to be one-third of the amount of data for 24-bit color image due to the depth map having only an 8-bit for depth information.

The results show that the visual quality of the reconstructed 3D image generated by the sparse multiview images of 4 × 4 images and the depth maps had a MOS of 4 or more when the

Fig. 6 Relationships between Mean Opinion Score and data size.

MOS of the 3D image reconstructed by the anchor is 5. In general, MOS of 3.5 is considered as an acceptable limit of visual quality deterioration. Figure 7(a) and (b) show reconstructed 3D images from 22×22 and 4×4 images, respectively. This indicates that even if the data amount of the 3D image is reduced to 4.3% by using this process, visual quality does not noticeably deteriorate in the reconstructed 3D image. On the other hand, in the 3D image generated from the 2 × 2 images and the depth maps, the quality deterioration from the 3D image reconstructed by the anchor became obvious, with a MOS of 2 or less.

Figure 7(c) shows the 3D image reconstructed from 2 × 2 images. The reason for the low score of MOS is caused by the part of the occlusion.

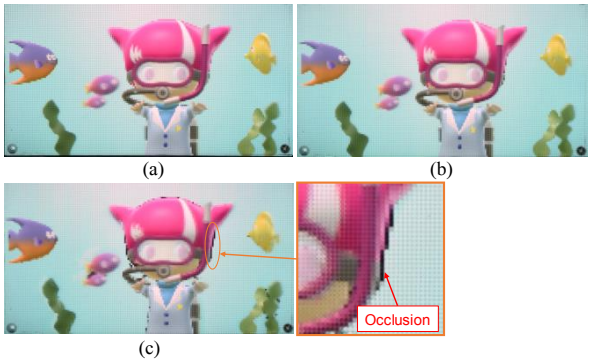

Fig. 7 3D image reconstructed from (a) 22 × 22, (b) 4 × 4, and (c) 2x2 images.

III. CODING EXPERIMENTS WITH THE REDUCED IMAGES

In this section, we encode fewer sparse multiview images and study the relationship between amount of compressed data and image quality for the reconstructed 3D images in each case. We apply the multiview video coding after converting from the elemental images to the multiview images instead of directly encoding the elemental images. The flowchart of compression for the multiview images in this experiment is shown in Fig.8.

We apply 3D high efficiency video coding (3D-HEVC) to the multiview images and study the influence of the coding to the reconstructed 3D image generated by multiview images in subjective evaluation experiment. We evaluate the influence from both reduction of the number of the dense multiview images and image degradation due to coding to the sparce multiview images.

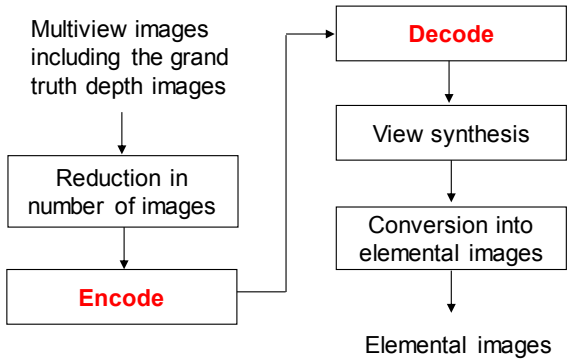

Fig. 8 Flowchart of compression of multiview images.

A. *Test sequences and equipment*

The previous results show that no significant deterioration in image quality was observed in the reconstructed 3D image even when the elemental images generated from sparse multiview images that were reduced to 4 × 4 images. Therefore, the elemental images generated from the sparse multiview images for the 4 × 4 and 8 × 8 images and the depth maps were used as evaluation images in this experiment. For comparison, the elemental images generated from dense multiview images 22 × 22 were used as an anchor. For the subjective evaluation experiment, the specifications of an integral 3D display and the elemental images are as shown in Table 1.

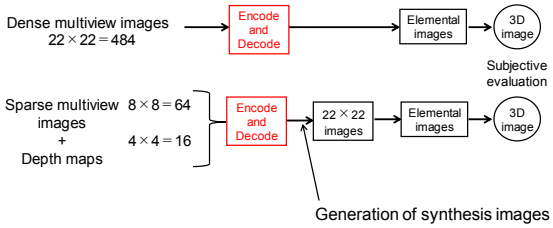

Fig. 9 Generation of anchor and evaluation images for subjective evaluation experiment.

B. *Experiment methodology*

3D-HEVC is used for multiview images with their depth maps. The maximum number of images that can be encoded at one time in conventional 3D-HEVC is 63 images. Therefore, we modified and extended reference software of 3D-HEVC so that up to 254 images could be encoded [19]. Moreover, to encode the dense multiview images of 22 × 22 images (484 images), the dense multiview images were divided into 4 groups.

Figure 9 shows the cases for the encoding/decoding for this subjective evaluation. The anchor image was based on the elemental images generated from the dense multiview images of 22 × 22 images that were not encoded or decoded. The depth maps were also encoded for the sparse multiview images of the 4 × 4 and 8 × 8 images since the depth maps are required for the view synthesis. To generate the reconstructed 3D images to evaluate each of the amount of compressed data, the sparse multiview images and depth maps were encoded with four quantization parameters (QPs) on the encoder. The view synthesis was rendered from the decoded sparse multiview images and the depth maps in the same way as in the previous section.

The synthesized images and the decoded sparse multiview images were converted into the elemental images, and the reconstructed 3D image was evaluated. For the subjective evaluation experiment, the DSIS method Variant II specified in ITU-R Recommendation BT.500 was used. The same as in the previous section, there were eight evaluators, and their position was 2.1 m from the lens array on the integral display.

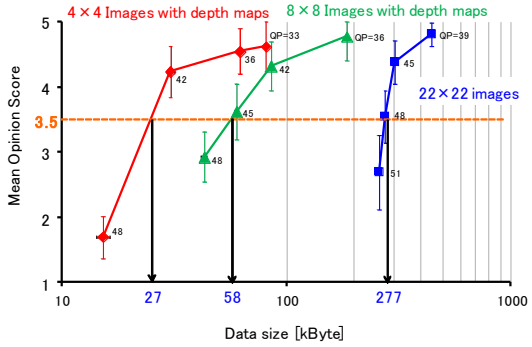

Fig. 10 Relationships between Mean Opinion Score and data size.

C. *Results*

Figure 10 shows the relationship between the amount of compressed data for each multiview image after encoding and the MOS of the 3D image displayed after decoding for each case. The data size along then x axis is the amount of compressed data after encoding. In the experimental results, the relationship between the amount of the compressed data and the visual quality of the 3D image is verified and compared between the cases with and without reducing the amount of the dense multiview images. The effect of visual quality degradation due to reducing the number of multiview images was not noticeable in the coding results. For this reason, highly efficient compression is achieved when the sparse multiview images and the depth maps for 4 × 4 images are coded. At this time, the amounts of data at which the MOS is 3.5 are 27 and 277 kBytes after encoding the multiview images of 4 × 4 and 22 × 22 images, respectively. Thus, the required amount of the data was reduced to about 1/10.

IV. CODING EXPERIMENTS WITH NATURAL CONTENT

The previous section showed the effectiveness of using view synthesis techniques to compress the integral 3D images in CG content. However, the ideal depth maps for view synthesis cannot be generated. Therefore, in this section, depth maps were generated from the actual multiview images captured by a camera array, and then the view synthesis was done to evaluate the compression performance for the integral 3D images. Figure 11 shows a flowchart of the image compression process. Compared with Fig.8, depth estimation was added before reducing the number of dense multiview images. To generate the depth maps and the synthesized images, we used Depth Estimation Reference Software (DERS) and View Synthesis Reference Software (VSRS) developed by MPEG [11,14]. Each piece of software was modified to correspond to the vertical parallax in addition to the horizontal parallax [20]. Furthermore, as in Section III, 3D-HEVC was used for coding experiments, and a subjective evaluation experiment was carried out.

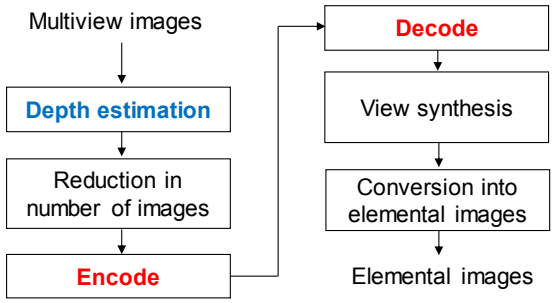

Fig. 11 Flowchart of image compression process.

A. Test images and equipment

We generated the test images acquired from a conventional 2D camera. The test images consisted of 625 (25 × 25) still images. To shoot the images, the camera was attached to an automatic x-y stage moving horizontally and vertically at 10 mm intervals. Figure 12 shows camera system to capture test images.

After being taken, the images were trimmed so that their resolution was 800 × 458 pixels. We photographed a doll and colorful flowers as shown in Fig.13. Considering the specifications of the 3D display system, the viewing distance from the evaluator to the display is 2.1 meters. For the subjective evaluation experiment, the specifications of the integral 3D display and elemental images are as shown in Table 1.

B. Depth map generation for experiment

DERS was used to generate depth maps corresponding to each acquired image. In the basic DERS algorithm, the target image and horizontal and adjacent images are used to generate depth maps. However, since the vertical parallax is

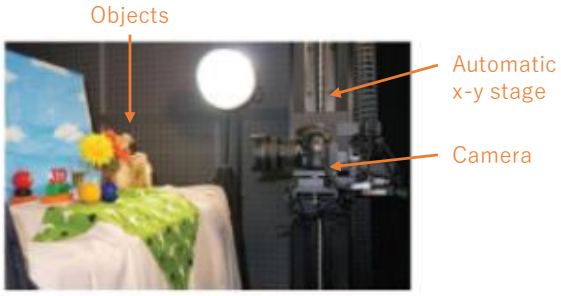

Fig. 12 Camera system to capture test images.

Fig. 13 Test images captured by the camera system.

included in the integral 3D image, the extended DERS corresponding to the vertical parallax is used. Figure 14 shows depth maps generated from two horizontal images and horizontal and vertical images and their respective methods for view synthesis. Image Fig.14(b) generated from four adjacent images is closer to the ground truth depth map than image Fig.14(a) generated from two adjacent images.

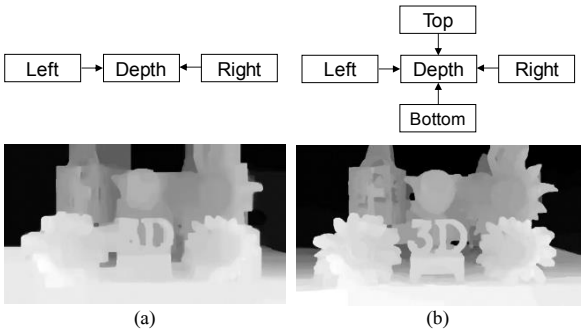

Fig. 14 Depth maps generated from (a) two horizontal images and (b) horizontal and vertical images.

C. Video coding and view synthesis for the experiment

3D-HEVC was used for compressing the multiview images that included the depth maps. Decoded images including the depth maps are used as information for rendering synthesis images by using VSRS. Figure 15 shows the workflow of VSRS. Unlike the view synthesis software used in Section III,

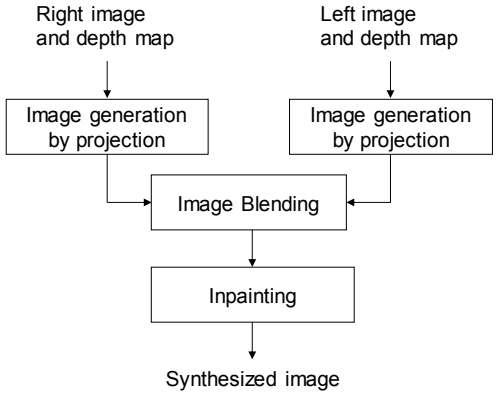

Fig. 15 Workflow of VSRS.

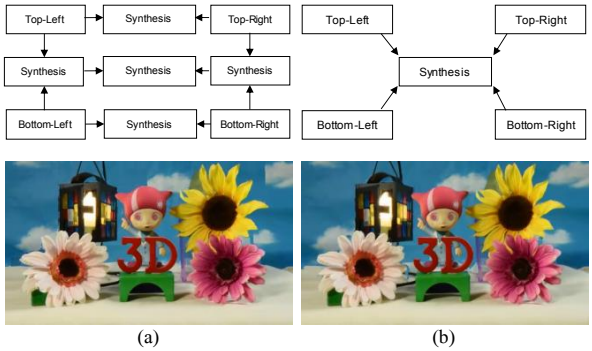

Fig. 16 Reference views for VSRS process. (a) Images synthesized by using (a) two and (b) four reference images.

VSRS fills holes that cannot be restored due to the influence of occlusion by using the inpaint algorithm [21].

In addition, the same as DERS, the VSRS uses four adjacent images corresponding to the vertical and horizontal parallax. Figure 16 shows the difference between images synthesized by using Fig.16(a) two and Fig.16(b) four reference images. In the case of image Fig.16(a), two steps had to be done to synthesize the target image. On the other hand, in the case of image Fig.16(b), view synthesis in the diagonal direction can also be rendered, so all the images can be synthesized in one step.

Figure 17 shows the results for peak signal-to-noise ratio (PSNR) between the four reference images and two reference images in a preliminary experiment. At 1.5 Mbit, the results of representing four reference images for both depth estimation and view synthesis has a PSNR about 3.4 dB higher than the results of representing two reference images. Furthermore, when only depth estimation is generated from four images, the results at 1.5 Mbit shows improvement of PSNR about 1.2 dB compare to the results of representing two reference images. Therefore, using four reference images seems to improve the accuracy of depth estimation and view synthesis.

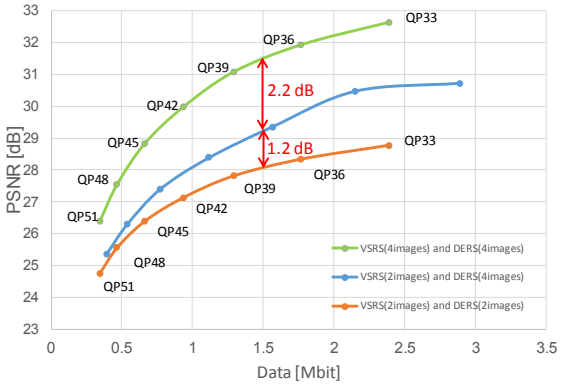

Fig. 17 Preliminary experiment for evaluating the difference in PSNR between four reference images.

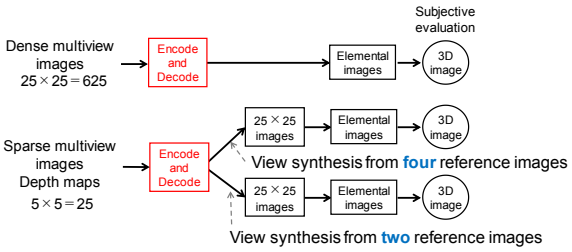

Fig. 18 Generation of anchor and evaluation images for subjective evaluation experiment.

D. *Subjective experiments*

To confirm the effect that using four reference images improves accuracy of the depth estimation and the view synthesis to reconstructed 3D images, we conducted a subjective evaluation experiment by displaying the integral 3D image. Elemental images were generated from the synthesized images and decoded sparse multiview images in each case, and a 3D image was displayed on the integral display.

We generated three cases for the experiments including anchor. The generation of anchor and evaluation images for the subjective evaluation experiments is shown in Fig. 18.

The anchor was based on the elemental images generated from the dense multiview images of 25 × 25 images that were encoded with four quantization parameters (QPs). To generate the evaluation images, sparse multiview images and depth maps were encoded, and the depth estimation and view synthesis were used by two cases of software from 2 reference images and from 4 reference images. In the subjective evaluation experiment, the DSIS method Variant II specified in ITU-R Recommendation BT.500 was used. The evaluator sat 2.1 m from the lens array on the display. Six evaluators participated in the experiment. Figure 19 shows the relationships between the MOS corresponding to each reconstructed 3D image for each QP and the data size for each

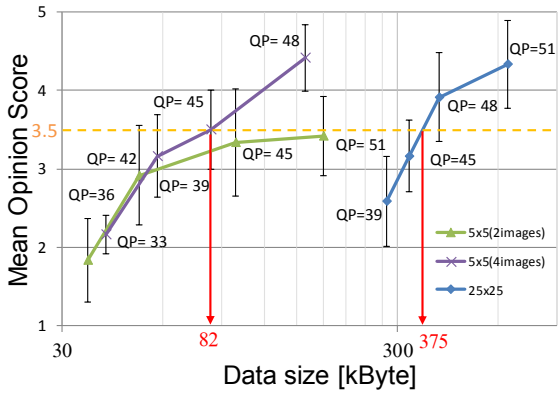

Fig. 19 Relationships between Mean Opinion Score and data size.

case. It is clear that the minimum data size for an acceptable 3D image reconstructed by 5 × 5 images from four reference images is 82 kBytes. This indicates that the coding performance in this case is quite efficient because the data size for a 3.5 MOS score for the anchor 3D image reconstructed by 25 × 25 images is approximately 375 kBytes. Thus, the required amount of the data for displaying an acceptable 3D image was reduced to about 2/9. Furthermore, the result of the view synthesis generated from four images is compared with the result of the view synthesis generated from two images. High efficiency compression is achieved for the images generated from four adjacent images the same as in the results for PSNR.

V. Conclusion

To advance toward 3D television services, we showed coding experiments on sparse multiview images and their depth maps after converting the sparse multiview images from dense multiview images to reconstruct an integral 3D image. The depth maps were generated from the dense multiview images before encoding, and reduced multiview images are generated during reconversion process from the sparse multiview images to the dense multiview images after decoding. The accuracy of depth estimation and view synthesis was improved by adding inter-view prediction for vertical direction and using four reference images rather than two images in the horizontal direction. In the results of a subjective evaluation to the natural test sequence, the reconstructed 3D images generated from the sparse multiview images and the depth maps achieved a 3.5 MOS with 2/9 as much data as the dense multiview images. In addition, it was shown that integral 3D images can be further compressed by improving depth estimation and view synthesis techniques.

Acknowledgment

The authors would like to thank Dr. Takanori Senoh of Tokyo Denki University, Tokyo, Japan, for his support in image processing the test images and M.Kano of NHK (Japan Broadcasting Corporation), Tokyo, Japan, for providing the test sequences in the experiments.

References

[1] F. Okano et al., "Real-time pickup method for a three-dimensional image based on integral photography," Appl. Opt. vol. 36, no. 7, pp. 1598-1603, 1997.
[2] H. Hoshino, F. Okano, and I. Yuyama, "Analysis of resolution limitation of integral photography," J. Opt. Soc. Am. A vol. 15, no. 8, pp. 2059-2065, 1998.
[3] J. Arai et al., "Integral three-dimensional television using a 2000-scanning-line video system," Appl. Opt. vol. 45, no. 8, pp. 1704-1712, 2006.
[4] J. Arai et al., "Integral three-dimensional television with video system using pixel-offset method," Opt. Express vol. 21, no. 3, pp. 3474-3485, 2013.
[5] P. M. G. Lippmann, " Épreuves réversibles donnant la sensation du relief", J. Phys., vol. 4, pp. 821-825, Nov. 1908
[6] X. Xiao, B. Javidi, M. Martinez-Corral, A. Stern, "Advances in three-dimensional integral imaging: Sensing display and applications", Opt., vol. 52, no. 4, pp. 546-560, Appl. 2013.
[7] J. Arai et al., "Progress Overview of Capturing Method for Integral 3-D Imaging Displays", Proc. IEEE, vol. 105, no. 5, pp. 837-849, May. 2017.
[8] Exploration Experiments for MPEG-I: Dense Light Field Compression, ISO/IEC JTC 1/SC 29/WG 11, Jan. 2018.
[9] R.N. Mekuria, K. Blom, P. Cesar, "Design Implementation and Evaluation of a Point Cloud Codec for 3D Tele-immersive Video", IEEE Transactions on Circuits and Systems for Video Technology, vol.27, no. 4 , pp828-842, 2016.
[10] K. Hara et al., "Coding Performance for Moving Picture of Integral Three-dimensional Image using 3D-HEVC", 3DSAp2/3Dp2-23L, pp. 1655-1656, 2016.
[11] "ISO/IEC JTC1/SC29/WG11 MPEG M36590", G. Lafruit, K. Wegner, T. Grajek, T. Senoh, K. P. Tamas, P. Goorts, L. Jorissen, B. Ceulemans, P. C. Lopez, S. G. Lobo, Q. Wang, J. Jung, M. Tanimoto, FTV software user guidelines, 2015.
[12] G. Tech, et al., "Overview of the multiview and 3D extensions of High Efficiency Video Coding", IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 1, pp. 35-49. Jan. 2016.
[13] K. Müller, P. Merkle, T. Wiegand, "3-D video representation using depth maps", Proceedings of the IEEE, vol. 99, no. 4, April 2011.
[14] Summary on MPEG-I Visual Activities on 6DoF, ISO/IEC JTC 1/SC 29/WG 11, Jan. 2018.
[15] "Methodology for the Subjective Assessment of the Quality of Television Pictures, ITU-R BT.500-13", Jan. 2012.
[16] N. Okaichi et al., "Integral 3D display using multiple LCD panels and multi-image combining optical system," Opt. Express vol.25, no. 3, pp. 2805-2817, 2017.
[17] H. Watanabe et al., "Wide viewing angle projection-type integral 3D display system with multiple UHD projectors", SD&A-358, pp. 67-73, 2017.
[18] Y. Iwadate, M. Katayama, "Generating Integral Image from 3D Object by Using Oblique Projection", 3Dp-1, pp. 269-272, 2011.
[19] K. Hara et al., "A method of increasing the number of views in HTM for the experiment of dense light field compression", ISO/IEC JTC1/SC29/WG11 Doc. M42150, Gwangju, South Korea, Jan. 2018.
[20] T. Senoh et al., "MPEG-I-Visual: Enhanced DERS for Quad Reference Views", ISO/IEC JTC1/SC29/WG11 Doc. M41955, Gwangju, South Korea, Jan. 2018.
[21] Tezuka Tomoyuki et al., "View synthesis using superpixel based inpainting capable of occlusion handling and hole filling", Picture Coding Symposium (PCS), pp. 124-128, May 2015.