A New Data Format for Multiview Video

MEHRDAD PANAHPOUR TEHRANI[†1] AKIO ISHIKAWA[†1]
MASASHIRO KAWAKITA[†1] NAOMI INOUE[†1] TOSHIAKI FUJII[†2]

This paper proposes a new data forma that can be used for multiview video representation. The new data format compensates the synthesis error at the viewpoint where the image is converted to a new format. In this scenario, residual based representations have already been proposed. In this paper we introduce and further discuss the new data format that we have recently proposed in our previous work. The converted image to the new data format is called hybrid image. NICT has developed a 200-inch 3D display. This display requires 200 HD multi-view images per frame at display side. We choose the hybrid representation as a novel data format for transmission of multi-view images to the display. A hybrid image consists of residual and reminder values, while the state-of-art use either residual or reminder. The representation based on hybrid image is a sequence of original and hybrid images that are repeated consequently in view/time direction. We evaluate the compression efficiency of represented multi-view images by hybrid data format. In this paper, we discuss a suitable camera configuration for our 3D display. Furthermore, we compare the compression efficiency of hybrid representation against other representations. Experiments demonstrate improvement of coding efficiency about 20% in average, using hybrid representation.

1. Introduction

3D video has attracted many applications such as 3DTV [1],[2] and FTV (Free-viewpoint TV) [3],[4]. By them, we can freely choose our desired viewpoint. However, current 3D display can only provide nearly one view angle at a moment. In order to realize true 3D, NICT has developed a 3D display, where user can see in 3D without glass, and has freedom to choose the view angle by moving horizontally in front of the display. NICT's display, realizes real life experience. The display is the world largest 3D multi-view display (200-inch) [5] that requires 200 full HD images per frame at display side.

To realize ultra-realistic communication, transmission of the 3D data is required. For auto-stereoscopic display, MPEG works on compression standard of view-plus-depth data representation [6], where the missing viewpoint can be easily synthesized and delivered to user before display.

Considering the large number of views, sparse multi-view-plus-depth (MVD) cannot be the best data representation for the system. Layered depth video (LDV) [7], and FTV data Unit (FDU) [8] were proposed based on MVD. They contain residual values. Compression of the residual causes loss of synthesis error, in area with small differences. Therefore, we proposed a hybrid representation [9].

We have previously proposed hybrid representation [9]. The hybrid representation consists of original and hybrid images that are repeated consecutively in spatial domain. A hybrid image has residual and reminder pixels [10]. Using the original views, we synthesize a virtual image, and estimate an error mask, at the same location of hybrid image. They are used for generation and reconstruction of the hybrid image. The regions corresponding to residual and reminder are distinguished by the estimated error mask. We compensate the virtual image by using hybrid image, in the reconstruction phase.

In [9], we experimentally demonstrated the effectiveness of our framework using several test sequences. The results showed that the reconstructed hybrid image has higher PSNR and fewer artifacts than the virtual image compensated by residual or reminder image. The raw data size of the hybrid images is also less than residual and reminder images.

In this paper, we focus on compression of multi-view images using hybrid representation. Regarding the compression, we introduce a suitable system configuration for NICT 200-inch 3D display. By using this configuration we lead to an optimal system configuration with respect to cost and view synthesis quality. For this configuration, we obtain better compression efficiency using our framework.

The rest of the paper is organized as follows. Section 2 is the introduction to NICT 200-inch 3D display. In section 3 we discuss the system architecture and compression configuration. The hybrid representation is briefly explained in section 4. Experimental results are shown in section 5. Conclusion of the paper comes in section 6.

2. The NICT 200-inch 3D Display

NICT has developed the world largest 3D multi-view display (200-inch) [5]. The display can be fed by 200 full HD images to compete with conventional HDTV or current 3D cinema. User can watch natural 3D without using special glasses. The viewpoint can be freely chosen by relocating in front of the display, as if we relocate in front of a real object.

The display consists of screen and a projector array. The screen is the combination of a diffuser next to a condenser lens. The diffuser narrows the angle of the incident light from behind in horizontal direction to provide accurate motion parallax, Figure 1, while the vertical diffusion is widen. The condenser lens converge the light rays to the designated viewpoint. There are 200 customized projectors that are aligned in slated matrix, behind the screen. Each projector has a unique horizontal location and provides a single viewpoint through the screen optics, with 33mm pitch. Optical centers are aligned perpendicular to the screen, see Figure 2. The viewing interval is 22.8mm.

†1 National Institute of Information and Communications Technology (NICT)
†2 Nagoya University

Figure 1 Tow views at a frame from two different angles show horizontal motion parallax.

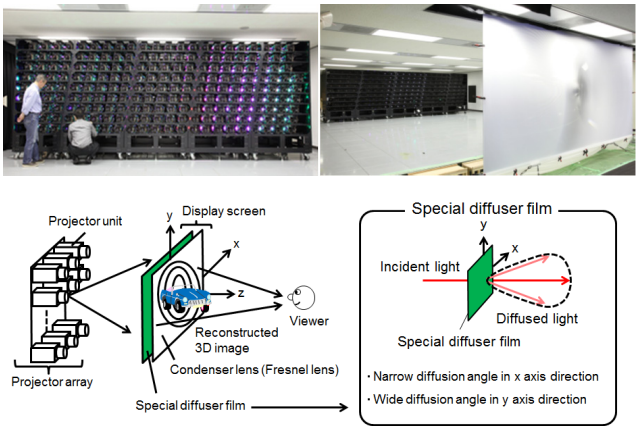

Figure 2 The NICT 200-inch 3D display, projector array and screen.

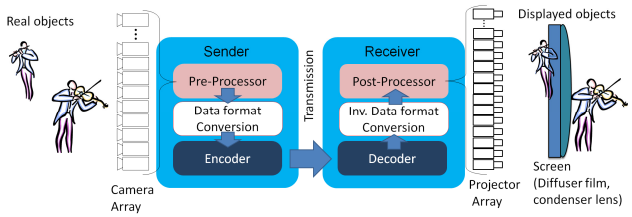

Figure 3 System block diagram for the 200-inch 3D display.

3. System Architecture

Figure 3 shows the overall system for the 200-inch multi-view 3D display. The system contains four main components, which are Camera Array, Sender, Receiver, and Projector Array.

Sender side contains two main units, which are pre-process, encoder. The pre-processor unit is responsible for geometry/color correction of the captured images by the camera array. Encoder has two components, which are data conversion and compression. The data conversion is where the hybrid images are generated. Compression is considered to be based on current standards, JPEG, MPEG-AVC (Advance Video Coding), or MVC (Multi-View Coding). Similarly, the inverse data conversion and decompression components are at the decoder unit of receiver side. The images after decoding are further corrected to be fed to the projector array in the post-processor unit.

3.1 Pre-experiment for Deciding the Number of Cameras

The system needs 1D convergent camera array as capturing system, similar to the characteristic of the projector array.

Number of cameras is a curial parameter that affects the quality of the displayed image, the processing complexity and construction cost. Subjective and objective pre-experiments are conducted to find suitable the number of cameras.

Figure 4 shows viewpoint 100 (view, estimated depth) of our test sequences. Each sequence contains 184 views, 22.8mm interval, full HD still images (1920x1080, RGB). They are very realistic, i.e. non-Lambertian reflectance surfaces.

Given these test sequences, we have estimated depth maps for several baselines at each viewpoint based on stereo-matching and graph cuts energy optimization technique [12]. The view synthesis is based on 3D warping [11] with depth maps.

Figure 5 shows an example graph for the abovementioned pre-experiment for a test sequence. In the graph of figure 7, the change in average PSNR over 150 synthesized images is shown versus the change in the baseline distance at depth estimation step. Considering PSNR more than 30~35 dB as acceptable quality, 4 baseline distances, i.e. 91.2mm, can be suggested as camera interval for the camera array. Further subjective evaluation also verifies that 4 baseline is suitable.

3.2 View Configuration

Based on the results obtained in previous section, we have chosen the architecture that is depicted in figure 6. For simplicity, the figure is illustrated for three views. In three view case, we need five views to generate the required data. It is due to the depth estimation process, where it used three views simultaneously to generate a depth map for the view in the center of the three views. Hence, for the 200-inch 3D display, we need 50+2 cameras for the camera array.

4. Multiview Video Representation Using Residual and Reminder (Hybrid)

Detail explanation for generation of a hybrid image is explained in our previous work [9]. However, we briefly explain the hybrid representation in this section.

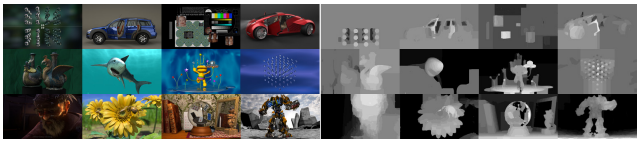

Figure 4 Snapshot of sequences, and the sequence number

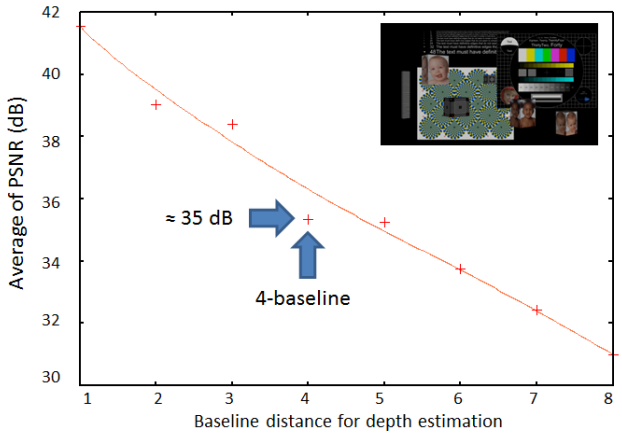

Figure 5 Average PSNR over 100 synthesized images versus the change in the baseline distance at depth estimation.

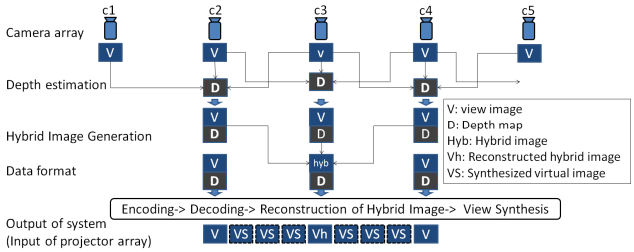

Figure 6 View Configuration for Compression.

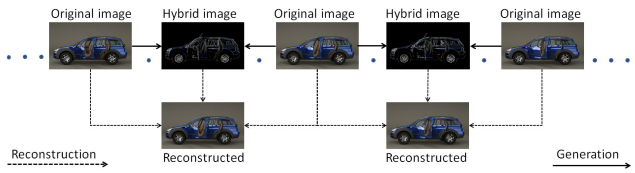

Figure 7 Configuration of views in hybrid representation.

As it mentioned earlier, hybrid representation consists of original and hybrid images repeated consecutively in spatial, i.e. view, direction. Figure 7 shows illustrated view configuration for hybrid representation. As shown in this figure, hybrid images are generated and constructed by using the neighbor images, i.e. left and right side. Figure 8 and Figure 9 summarize the generation and reconstruction of a hybrid image that will be discussed in the following.

Note that it also possible to generate and reconstruct a hybrid image using a single view that is located near the target image. Target image is converted to hybrid image.

4.1 Generation of Hybrid Image

Since our representation is an alternative data format for MVD, we explain the procedure given the left and right view-plus-depth of *VT*, i.e. *VL/DL* and *VR/DR*.

An image represented in hybrid format, consists of residual and reminder values. Figure 2 shows example of hybrid image, residual image, reminder image, estimated error mask for synthesized image at the location of the hybrid image, and the original image. The generation algorithm is as follow:

- View Synthesis: Intermediate virtual view at target view location is synthesized by using two reference views and their depth maps at left and right sides, i.e. VL/DL and VR/DR. The view synthesis is based on 3D warping [11] with depth maps.
- Estimation of Synthesis Error Mask: Figure 10 depict the procedure for estimation of synthesis error mask. Note that we call it “ESTIMATED” since the procedure is simply NOT the subtraction of original image and synthesized image, followed by throsholding.
- Residual Image Generator: Subtraction of a virtually generated and the original target images. It is followed by 1-bit reduction in bit-plane depth of the subtraction result, which is the outcome of a mapping process.
- Reminder Image Generator: Reminder image is the output of modulo operation on each pixel value of target image given a divisor (D) [10]. D is chosen from a look up table (LUT), given a gradient value of the same location. Gradient values are pixel values in gradient-like image that is generated from the synthesized virtual image at the target view location. Detail of gradient-like image generation is explained in [9].
- Hybrid Image Generator: In error mask, the areas with intensity of “0” are the area with low error value, so the hybrid image in these areas is represented by reminder values. The rest of areas correspond to high synthesized error; therefore they are represented by residual values in the hybrid image.

4.2 Reconstruction of Hybrid Image

The generation algorithm is simply as follow:

- View Synthesis: Similar to synthesis at generation phase using VL/DL and VR/DR.
- Estimation of Synthesis Error Mask: Similar to mask estimation at generation phase using VL/DL and VR/DR.
- Reconstruction of Hybrid image (Residual and Reminder): Using the estimated error mask at step 2, we distinguish the reminder and residual areas. Then we apply the inverse of generation processes for residual and reminder parts [9], [10], respectively.

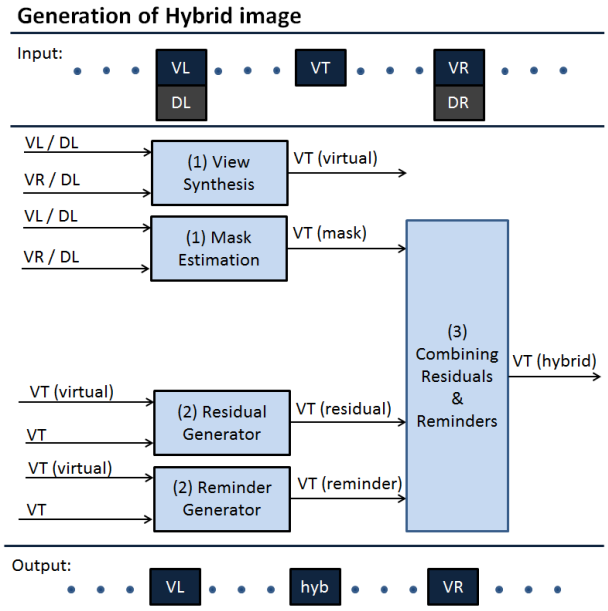

Figure 8 Procedure for generation of a hybrid image.

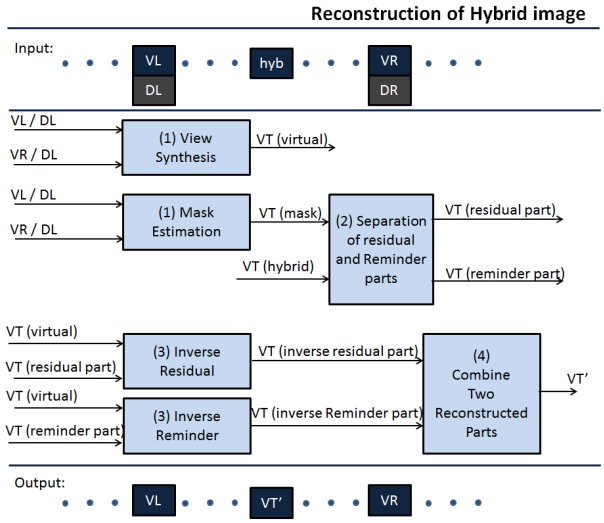

Figure 9 Procedure for reconstruction of a hybrid image.

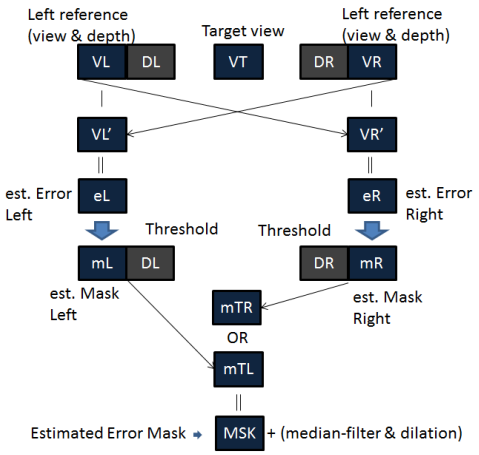

Figure 10 Procedure to estimate synthesis error mask.

5. Experiments

We used the configuration discussed in section 3 for the experiments.

Based on configuration of figure 6, we have conducted the experiment for a camera array that has 9 cameras, 1D convergent. The camera interval is 4-baseline, i.e. 91.2mm. The camera configuration through the experiment is illustrated in figure 11. Depth map at location of c2, to c8 are estimated using stereo matching and graph cuts optimization. c0 and c9 are only used for depth estimation at c2 and c8 at sender side, respectively. The camera views that are drawn between each pair from c2 to c8 are needed to be synthesized and delivered to the projector array. Views at c3, c5, and c7 are used through the experiments to evaluate the compression performance.

The comparison efficiency is measure by the rate distortion curves for unconverted images at (c3 c5 c7) and the converted image in the form residual, or hybrid at (c3 c5 c7.) Intra coding is applied in the experiments. The average decoded PSNR and the total bit rate of (c3 c5 c7) are measured.

In order to compress a hybrid image, a local decoder is placed at encoder side. We firstly compress and decompress *VL/DL* and *VR/DR*, respectively. Then, the data format conversion step is applied, followed by the compression of the new data format.

The compression ratio is adjusted by vary the QP, while it is the same for all images, i.e. image, depth map, hybrid image, residual image.

Parameters for hybrid generation/reconstruction are as follows. Threshold value for estimation of the error mask is set to 4. *D* value for reminder part of the hybrid image is 8 for all areas when the gradient value is more than 4. The rest of areas are filled with synthesized image at decoder.

Figure 12 shows RD curves for three test sequences. According to the graphs, in rates from low to high the hybrid representation outperforms, and in very high bit rate residual representation is the best. Similar behavior was observed for all of test sequences.

Subjective assessment demonstrates similar quality for the residual and hybrid. However, given the coding condition, i.e. not well-optimized, and the configuration we assume in the paper, i.e. figure 6 and 11, the coding efficiency using hybrid images is averagely improved about 20% in comparison with residual representation.

We expect better performance if QPs are carefully assigned, given a target bit rate. Note that the best combination rate of reminder and residual pixels in a hybrid image is dependent on the baseline distance, accuracy of depth estimation and view synthesis. If we use better optimization techniques for depth estimation [13], e.g. belief propagation, and view synthesis using graph cuts optimization with reliability reasoning [14], we can further improve the coding efficiency for hybrid representation.

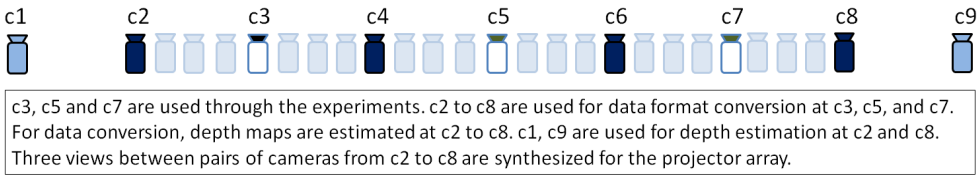

Figure11 The camera configuration through the experiments.

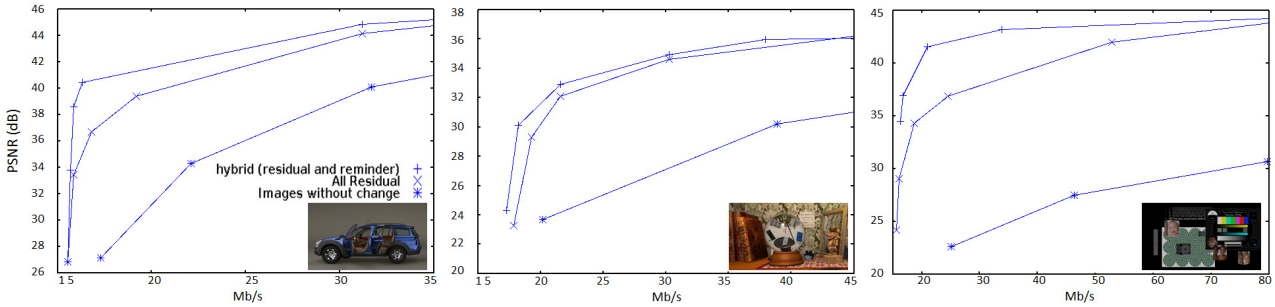

Figure12 Average decoded PSNR of three views versus total bit rate of the three views. The three views are (c3, c5, c7) of Figure 11.

6. Concluding Remarks

We have proposed hybrid image as a novel data format that can compensate the synthesis error in the multiview video coding better than conventional formats, such as residual or reminder images. In this paper, we have demonstrated the coding performance of hybrid representation on multi-view images. In addition, we found a suitable configuration for transmission of the multi-view images to the NICT 200-inch 3D display where the number of cameras is minimized and the acceptable subjective and objective qualities were maintained. In the experiments, we showed better coding efficiency of the hybrid representation against other representations. The coding configuration and view architecture for the data format is important for achieving the highest performance. The parameters through the generation of hybrid image are also curial in the achieving the optimal coding efficiency.

Hybrid image contains interview redundancy, as it is also available in residual and reminder images. Therefore, our future direction is to evaluate coding performance of hybrid representation using multi-view coding (MVC) on the computer generated test sequences, as well as actually captured multi-view videos by a camera array. Parameters and coding configuration will also be optimized for improving the coding efficiency.

Reference

1) A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, C. Zhang, “Multiview imaging and 3DTV”, IEEE Signal Processing Magazine, 24(6):10-21 (2007).
2) A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Muller, M. Lang, “Three-Dimensional video postproduction and processing”, in Proc. IEEE 99(4), 607-625 (2011).
3) M. Tanimoto, M. Panahpour Tehrani, T. Fujii, T. Yendo, “Free viewpoint TV,” IEEE Signal Processing Magazine 28(1), 67–76 (2011).
4) M. Tanimoto, M. Panahpour Tehrani, T. Fujii, T. Yendo, “FTV for 3D spatial communications,” Proc. IEEE, 100(4), 905-917 (2012).
5) S. Iwasawa, M. Kawakita, “Qualifying capabilities of the prototype 200-inch automultiscopic display,” in Proc. 3DSA, 2011, S1-15, 105-109 (Jun. 2012).
6) C. Fehn, “Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV”, in Proc. SPIE Stereoscopic Displays and Virtual Reality Systems, XI, pp.93-104 (Jan.2004).
7) K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff T. Wiegand, “Reliability-based generation and view synthesis in layered depth video“, in Proc. IEEE Intl Conf on MMSP, 34-39 (Oct.2008).
8) M. Tanimoto, M. Wildeboer, “Frameworks for FTV coding”, in Proc. PCS, 1-4 (May 2009).
9) M. Panahpour Tehrani, A. Ishikawa, M. Kawakita, N. Inoue, T. Fujii, “A hybrid represenation for multi-view image,” in Proc. 3DTV-CON 2012, (Oct. 2012).
10) M. Panahpour Tehrani, T. Fujii, M. Tanimoto, “The adaptive distributed source coding of multi-view images in camera sensor networks”, IEICE Trans, E88-A(10) (2005).
11) Y. Mori, N. Fukushima, T. Yendo, T. Fujii M. Tanimoto, “View generation with 3D warping using depth information for FTV,” Signal Proc Image Com., 24, 65– 72 (Jan. 2009).
12) Y. Boykov, O. Veksler and R. Zabih, “Fast approximate energy minimization via graph cuts,” IEEE Trans. Pattern Anal. Machine Intell., 23, 1222–1239 (Nov. 2001).
13) R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M.Tappen and C. Rother, “A comparative study of energy minimization methods for Markov random fields with smoothness-based priors,” IEEE Trans. Pattern Anal. Machine Intell, 30(6), 1068-1080 (2008).
14) L. Yang, T. Yendo, M. Panahpour Tehrani, T. Fujii and M. Tanimoto, “Probabilistic reliability based view synthesis for FTV”, in Proc. ICIP, 1785-1788 (Sep. 2010).