3-7 MPEG Multi-View Image Coding Standardization

SENOH Takanori, YAMAMOTO Kenji, OI Ryutaro, and KURITA Taiichiro

Standardization of multi-view image coding, which is the seed of 3-dimensional Videos or free-viewpoint TVs is going on in the Moving Picture Expert Group (MPEG) under International Organization for Standardization (ISO). MVC (Multi-view Video Coding) compresses the multi-view images by view prediction estimating disparities between views as well as frame prediction estimating the motion vectors between frames. Another new movement is an investigation on a higher coding efficiency by estimating the image depths. In this paper, these multi-view image coding technologies are introduced together with the research performed in NICT.

Keywords
3-dimensional video, Free-viewpoint TV, Multi-view image, Disparity-compensated prediction, Depth estimation

1 Forward

Ultra-realistic communication is being researched in order to raise the level of the means of communications between people[1]. The aim of ultra-realistic communications is the development of highly-realistic communications using deep three dimensional images.

The method of producing such three dimensional images consists of sending separate images to the left and right eyes in a stereoscopic image format, expanding these images in a multi-view image format reproducing three dimensional object image space in a volumetric format and reproducing the intensity and phase of the light emitted from the three dimensional objects in an holography format[2]. Each of these methods utilizes multi-view images captured from multiple differing viewpoints.

Although multi-view images have a wide range of usage such as producing free-viewpoint television (FTV)[3] that enables users to view objects from a free viewpoint, it is necessary to transmit multi-view images that put a large burden on transmission and storage.

In order to compress multi-view images used for free viewpoint images and three dimensional images, the Moving Picture Expert Group (MPEG) under the International Organization for Standardization (ISO) has been standardiszing a method for multi-view image coding from 2001 and published MVC (Multi-view Video Coding) as Annex H (Multiview video coding) of MEPG-4 Video Part10 (Advanced Video Coding) coding standardization (ISO/IEC14496-10/ITU-T H.264) used in TV broadcasts for mobile phones and high-density optical disks, etc., in 2009 for FTV's phase 1[4].

This method utilizes the motion compensated inter-frame prediction of conventional MPEG-4 AVC (the same method as ITU-TH.264) for the disparity-compensated prediction between viewpoint images, and been used as a 3D image coding method for high-density optical disks. In addition, as a FTV's phase 2, standadization started in 2007 on the 3DV/FTV (3-Dimensional Video/Free-viewpoint TV) coding for the purpose of further improving coding efficiency using image depth information[5].

In the next chapter, we will explain the technological content of these coding standards and discuss an adaptive depth estimation examined at NICT.

2 Multi-view Video Coding (MVC)

2.1 Multi-view images

As shown in Fig. 1, multi-view images are captured by multiple cameras for one scene and the number of images produced from different viewpoints depends on the number of cameras. If the distance between the cameras is the same as the distance between the human eyes (approximately 6.5cm), the random two camera images can be used as a stereoscopic image. In regard to stereoscopic images, such devices are put into practical use as polarized glasses or shutter glasses that send separate images to both the left and right eyes, and a glasses-free stereo display enabling each left and right eye to view different images owing to a disparity barrier. In addition, by using a larger number of multiple camera images, such images can be used for glasses-free 3D images which continuously display the viewing zone of each image.

2.2 Camera alignment and test images

As shown in Fig. 1, high correlation can be observed as the same objects appear in the multi-view images acquired from each camera. Consequently, the Multi-view Video Coding (MVC) was standardized using the correlation between viewpoint images. When performing standardization, the alignment of each type of camera was proposed as shown in Fig. 2 and each of the test images was provided by the participants in the standardization. Initially, it was proposed that the rectification which corrects the alignment of each camera by the projective transformation of images be conducted during coding. However, since the codec would have incurred a large burden, correction was made before coding.

For this correction, after correcting camera

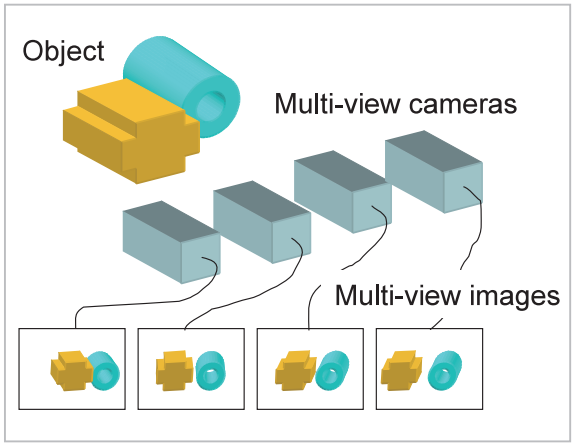


Fig.1 *Capturing of multi-view image*

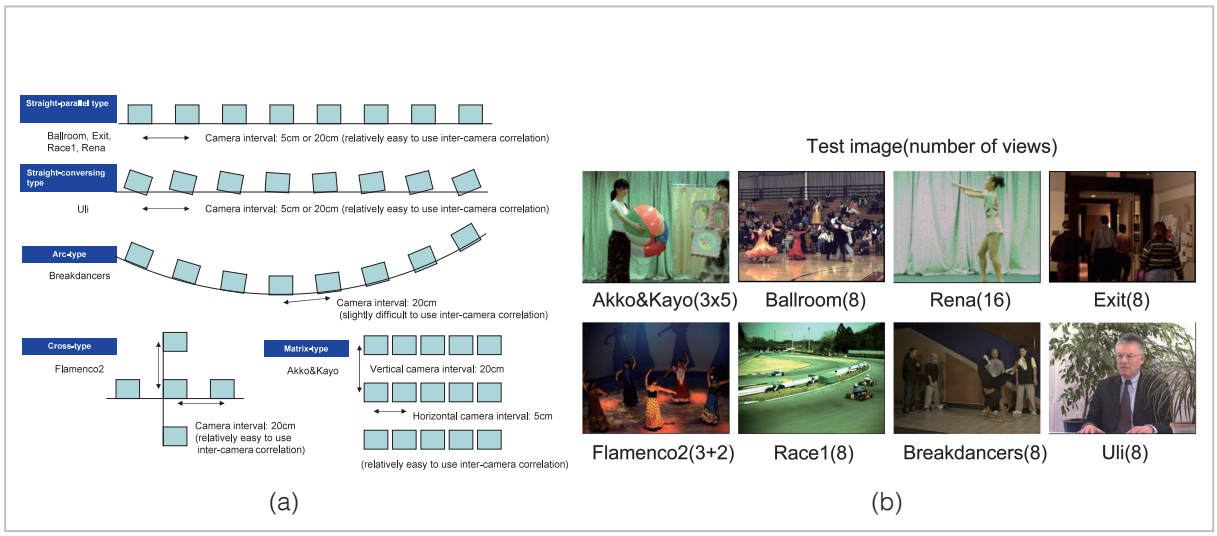


Fig.2 *Camera alignment and test images*

(a) Camera alignment example, (b) test image example

lens distortions by capturing given plaid patterns, the object points that can be deemed vanishing points and infinitive distances of the plaid patterns in the images are specified as the corresponding points of each viewpoint image, a projective transformation matrix of each viewpoint image is acquired in the way the positions of these infinite-points can coincide between all the viewpoint images, and the projective transformation is applied to each viewpoint

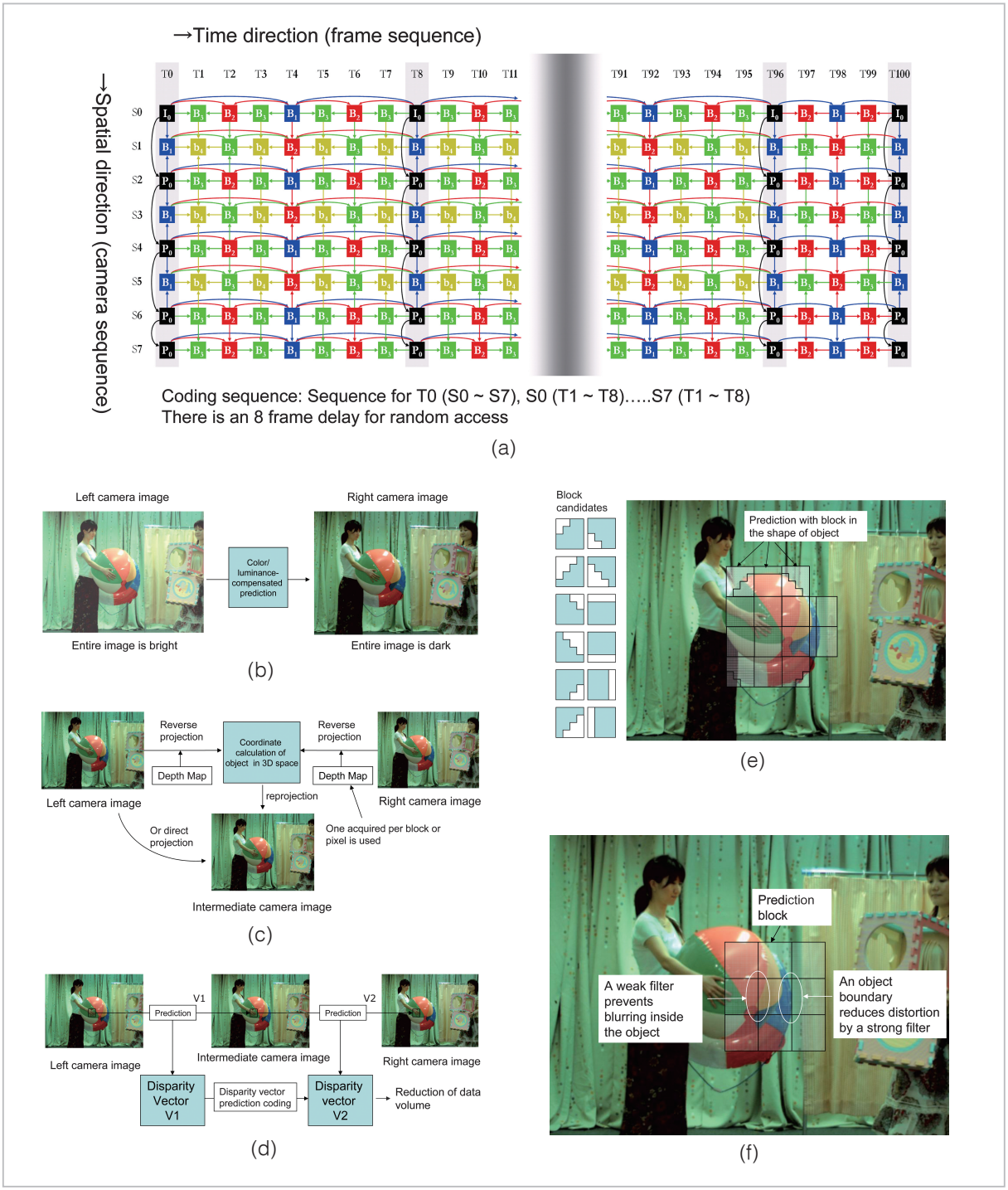


Fig.3 *Each type of coding methods proposed for the standardization of multi-view image coding (MPEG-4 MVC)*

(a) Spatio-temporal motion compensated prediction, (b) Illumination-compensated prediction, (c) Projective transformation prediction, (d) Disparity vector prediction, (e) Asymmetric macro block prediction, (f) Adaptive prediction filter

image. This enables all the camera directions and internal parameters to become aligned. Next, a common projective transformation matrix is acquired in the way the epipolar lines passing through the corresponding points in each viewpoint of a certain feature point within a finite distance can become parallel, and once again the projective transformation is applied[6].

2.3 Proposed coding method

All types of coding methods were proposed using these test images. The main details are shown in Fig. 3. Figure 3(a) shows the spatio-temporal hierarchical bi-directional prediction. There is an existing method used for the inter-frame prediction that the direction and distance of the block having moved between frames is acquired by motion estimation with block matching, and the block is predicted by applying the estimated motion vector to a block in another farme, then, transform coding, quantization and variable-length coding is conducted for the remaining difference to send. The spatio-temporal hierarchical bi-directional prediction is adopted by this method, which applies this motion vector estimation between images at different viewpoints, conducts the parallel translation search for block between viewpoint images, adds the disparity-compensated block with the acquired vector to the inter-frame prediction candidate, and predicts the multi-view images both from between frames and from between viewpoints. Between differing viewpoint images, the angle from which the object can be seen is different, there is a problem with the occlusion that the foreground hides background not to be seen, and therefore the prediction efficiency is not very high. However, the coding efficiency is raised by using hierarchical bi-prediction for prediction between frames of high correlation.

Figure 3(b) shows the illumination-compensated prediction method that raises the prediction coding efficiency by adding the value compensating luminance and chrominance difference between viewpoint images due to lighting position and the color sensitivity dispersion of cameras as the offset obtained from the average value between viewpoint images. This color matching between viewpoint images was not adopted since the codec burden was reduced by conducting color matching on the camera side.

Figure 3(c) shows the projective transformation prediction method that acquires the disparity (image depth) through the corresponding point matching between viewpoints images and then predicts the viewpoint images by conducting projective transformation using the acquired disparity. There were methods of projective transformation: one that acquires the 3D coordinate of the object from the disparity and camera parameters, and then conducts the prediction by projecting the acquired 3D object onto the viewpoint image that one wants to predict, and one that divides the predicted viewpoint images into several blocks, approximating that each block is a planar projection in the 3D space, and conducts prediction by acquiring the homography matrix between viewpoint images that is the generalized affine transformation. However, these were not adopted because of less prediction efficiency for more complex processing.

Figure 3(d) is a method that reduces the disparity vector data volume by conducting prediction coding for disparity vectors between viewpoint images. However, since the disparity vector data volume was small in comparison to the image data volume and accordingly the coding efficiency did not improve to a large extent, this method was not adopted.

Figure 3(e) is an asymmetric macro block prediction method. This method increases the accuracy of the block predictions by selecting the block shape in line with the shape of each object with a different disparity volume. However, since the prediction efficiency did not improve due to the increase of information used to select blocks, this method was not adopted.

Figure 3(f) is an adaptive prediction filter method. Since images become discontinuous along with block distortion when performing disparity-compensated prediction at the block unit level and then resulting in large disparity at

the block boundary, this method operates smoothing by adaptively applying the low-pass filter only to the parts with large disparity differences. Although this method subjectively reduces distortion, it was not adopted due to images becoming blurred.

Following this, a method which predicts the movement vectors of the neighboring multi-view images using disparity vectors was proposed. However, this was not adopted for the reason that the improvement of coding efficiency was low at approximately 0.5 dB.

2.4 Standardized coding methods

Ultimately, the spatio-temporal hierarchical bi-directional prediction method shown in Fig. 3(a) was adopted to standardize the Multi-view Video Coding (MVC). This method raises the coding efficiency by further layering the bi-directional frame distance prediction (B frame in the figure) that enables the inter-frame difference between the frames of still objects to become 0 and results in the large compression effect. Conversely, in regard to the prediction of inter-viewpoint images, since disparities always occur even if objects are still, the disparity-compensated prediction of the entire images becomes essential. When disparities occur, the background that hid behind the foreground in the neighboring viewpoint image becomes visible, and since the viewing angle changes as the viewpoint changes of what was rather visible in the neighboring viewpoint image, the corresponding block shape distorts. Therefore, the prediction by simple parallel translation of pixel block does not raise prediction efficiency and large compression ratio cannot be expected. Although the improvement of coding efficiency by MVC was approximately half as much as that when coding and transmitting the entire viewpoint images by the existing MPEG-4AVC, MVC was adopted as the 3D image coding method for high-density optical disks.

3 3D image/free-viewpoint image coding (3DV/FTV)

From 2007, MPEG commenced the examination of a 3DV/FTV coding method as Phase 2 of the multi-view image coding in order to improve the coding efficiency. As shown in Fig. 4, the method reduces the data volume by coding a small number of viewpoint images and their depth information (depth map), performing transformation based on the depth information on the decoding side and synthesizing the required viewpoint images, as opposed to coding all multi-view images.

This method will be used for coding 3D images based on the multi-view images and free-viewpoint video services that enable viewers to freely view scenes from the position they desire since random multi-view images can be highly accurately synthesized if the cameras are accurately aligned, color sensitivity is accurately uniformed and the depth information is accurate. Furthermore, this method is expected to be used as a holographic image coding method in the future since, as shown in Fig. 5, by using depth information, the wave front of light emitted from objects can be calculated

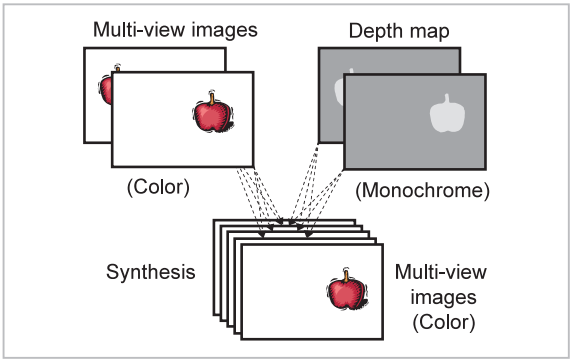


Fig.4 *Synthesis of multi-view images by depth maps*

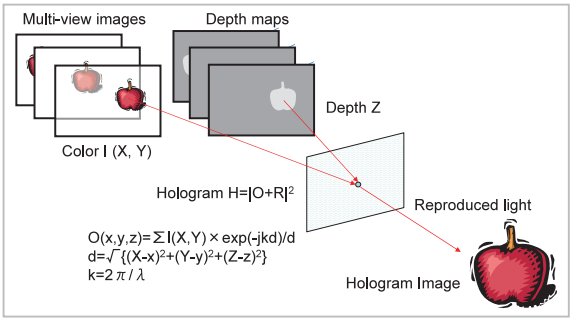


Fig.5 *Electronic holography reproduction from depth images*

and by using such calculations, ideal 3D images can be achieved[7]. In order to achieve these applications, it is necessary to extract depth from multi-view images.

3.1 Depth of multi-view images

The relationship between multi-view images and their depth is shown in Fig. 6. Figure 6 shows the case where the cameras with the same focal length f are aligned on baseline X in the equal distance L in horizontal and parallel positions. Following formula represents the relationship between distance D from baseline X to a point P on the object and the corresponding pixel positions x_1, x_2 in the respective camera images.

$$\frac{f}{D} = \frac{x_1 - x_2}{L} \tag{1}$$

While the height of the point in the figure is set as Y=0 for the purpose of brevity, formula (1) holds for any value of Y. The gap $(x_1 - x_2)$ between the positions of the corresponding pixel positions is called disparity and is often used as the depth values of multi-view images. When representing the depth of object by the disparity, there are such advantages as shifting a camera image pixel by its disparity enables for easily synthesizing the neighboring camera image and the disparity can be used as the depth value of image when displayed as a 3D image. In the following, the value represented by the disparity is used as depth value. In order to keep the accuracy of depth value, the depth value is to be normalized so that the depth value can be represented in full scale. Given the maximum disparity d_{max} and minimum disparity d_{min} per scene, normalization is conducted by the following formula when representing the disparity in 8-bit (0 ～ 255),

$$depth = \frac{255(d - d_{min})}{d_{max} - d_{min}} \tag{2}$$

3.2 Depth Estimation

3.2.1 Corresponding point matching

In order to obtain the depth from multi-view images, first, corresponding point matching is conducted and the pixel position displacement to make matching error minimal is acquired. Specifically as is show in Fig. 7, pixel position $pix\ (x, y)$ of camera 1 image (View 1) is gradually shifted (d=0, 1, 2, …) to all pixels $pix\ (x, y)$ of camera 2 image (View 2) and displacement d to make the absolute difference of pixel value $|pix\ (x+d, y) - pix\ (x, y)|$ minimal is searched.

$$D = \left| pix(x + d, y) - pix(x, y) \right|$$
$$d = \min_d \{D\} \tag{3}$$

Since this corresponding point matching is vulnerable to camera gaps and noises, in order to raise reliability, the color difference value is used as pixel value in addition to the luminance value, and 3×3 pixel block matching is performed.

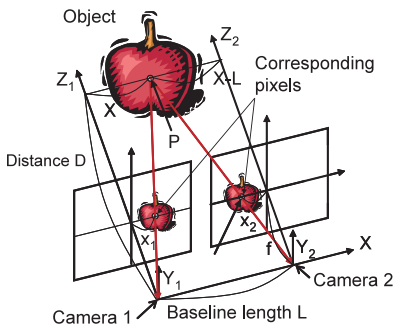


Fig.6 *Relationship between multi-view images and depth*

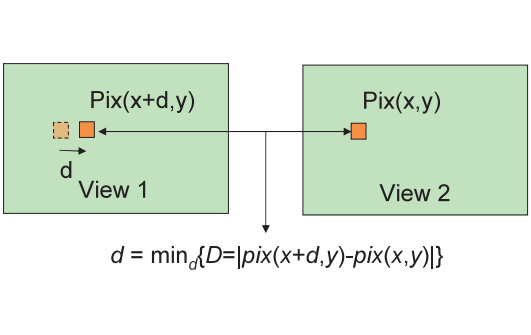


Fig.7 *Corresponding point matching*

3.2.2 Smoothing

The corresponding point matching is susceptible to the effects of errors in camera image rectification, color matching errors between camera images, noise mixing in images[8]. Consequently, the acquired depth map contains many errors. For this problem, using Belief Propagation and Graph Cuts theory, smoothing of estimated depth value is conducted so as to minimize the evaluation value that is the addition of the weighted difference from the depth value in the neighboring pixels (referred to as continuous constraint of depth value) to matching error[9][10].

$$E = \sum_{x,y} \{D + w|d(x+1,y) - d(x,y)|\}$$
$$d = \min_d(E) \quad (4)$$

In this smoothing process, as is shown in Fig. 8, from the matching errors to all the depth value candidates in all the pixels acquired by corresponding point matching, considering discontinuity of the depth value on the object boundary, the depth value is determined in the way that the evaluation value in the entire image becomes minimal (local minima in reality).

Although on individual objects, the sum of evaluation values can be decreased by equaling depth values as much as possible, on the object boundary, depth value is discontinuous and evaluation value increases. For that reason, by detecting the object boundary according to color difference (segmentation), the reduction of the weight of depth continuity is also performed. However, if specifying the object boundary only by the difference of colors, different colors of a same object induce the depth gap to result in the lower quality of depth map.

Furthermore, in order to reduce flickers in the temporal direction, still parts are detected by inter-frame pixel comparison and then arithmetic mean of matching error D derived in previous frames is calculated only for those parts, or after acquiring manually the depth value of still parts and making matching error D at those depth values 0, Belief Propagation and Graph Cuts process are applied.

3.2.3 Occlusion and pseudo-matching

In addition to camera gaps and noises, as is shown in Fig. 9, the occlusion issue that the corresponding points cannot be found hiding behind the foreground object and the issue of pseudo-matching in the uniform texture[11][12].

While the occlusion issue can be reduced by selecting the minimum value from matching errors to multiple camera images, objects with repetitive patterns or uniform textures readily generate the pseudo-matching at erroneous depth values. There is a method for reducing this issue using the average value for matching errors per multiple camera images. However, there is the problem that matching errors contain high noise in the occlusion parts and makes it difficult to estimate the optimum depth value.

3.2.4 Adaptive depth estimation

We will now discuss the adaptive depth estimation examined at NICT below. This adaptive depth estimation adaptively switches between the minimum value and average value of multiple matching errors and reduces the miss-estimation both in occlusion and pseudo-matching. Here, in order not to increase the calculation amount of depth estimation, as shown in Fig. 10, the images captured by 3 cameras aligned in horizontal and parallel positions are used. Varying depth value d from 0 to the maximum

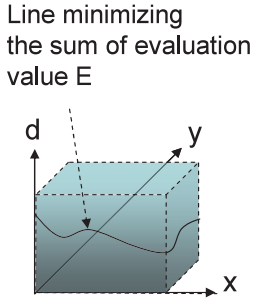


Fig.8 Smoothing

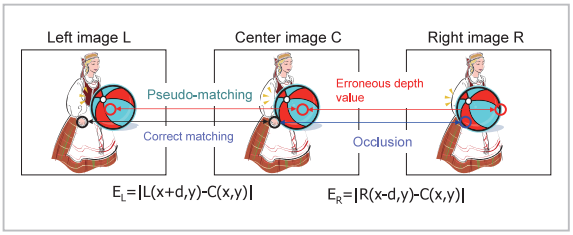


Fig.9 Occlusion and pseudo-matching

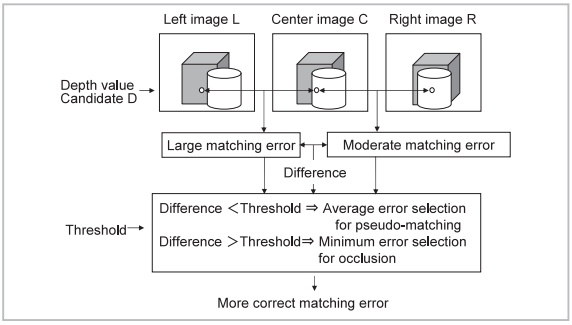


Fig.10 *Adaptive depth estimation*

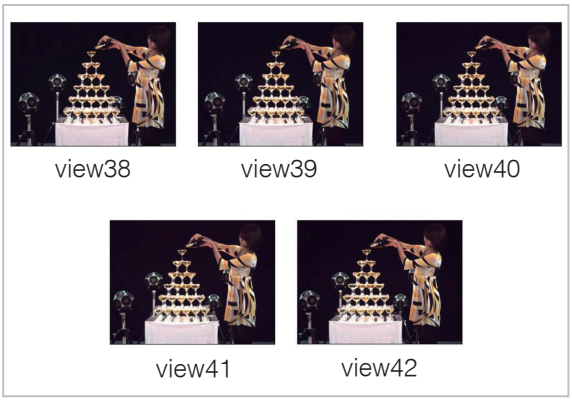


Fig.11 *Test images (Champagne Tower)*

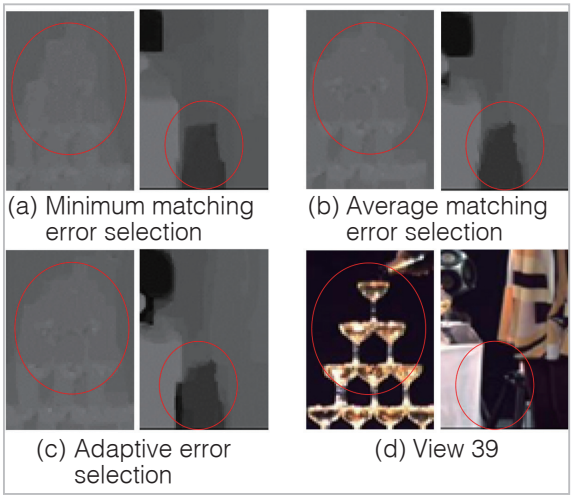


Fig.12 *Depth maps and View 39*

disparity per pixel of center camera image C, absolute difference D_L, D_R between the corresponding pixel values in neighboring camera images L, R and the pixel value in the center image are derived for the left and right each. When the difference between the left and right matching errors is larger than threshold value, it is judged that there is a high possibility of occlusion being occurring in one image and the depth value is estimated using the less matching error. When the difference is less, judging that occlusion is not occurring, in order to reduce the possibility of pseudo-matching, the depth value is estimated from the average value of the left and right matching errors. The threshold is determined by an object on the empirical basis, the optimum value is approximately one-tenth of the maximum matching error.

In the following, the effect of adaptive matching error selection is shown. A part of the test image (Champagne Tower), shown in Fig. 11, provided by Nagoya University that has been used for 3 DV/FTV standardization study was used for the experiment[13]. Camera interval is 5cm and rectification and chromatic compensation has been already applied to the image.

In the experiment, the depth map of View 39 is first acquired based on the algorithm shown in Fig. 10, using View 38, 39, 40 and likewise the depth map of View 41 is acquired from View 40, 41, 42, and then intermediate View 40' is synthesized from these 2 depth maps and compared with camera image View 40. Taking externally-given parameters as the selection criterion for the average value and minimum value of matching errors, th = 33 was used through a preliminary experiment in the case of the maximum matching error = 255. For depth estimation, the reference software (DERS5) that has being developed in MPEG/3DV group was used with some adjustment[14]. The software before adjustment acquires the stereo matching error in the left and right views at all the provisional depth values using 3 view images, and after selecting a depth candidate which is giving the minimum matching error, determines the depth value of each pixel applying the smoothing process based on the Graph Cuts theory. Figure 12 shows the depth maps when selecting the minimum matching error with the adjusted software, when selecting the average value for matching errors, and when performing adaptive matching error selection that conducts the minimum matching error selection at th $\geqq$ 33 and otherwise the average

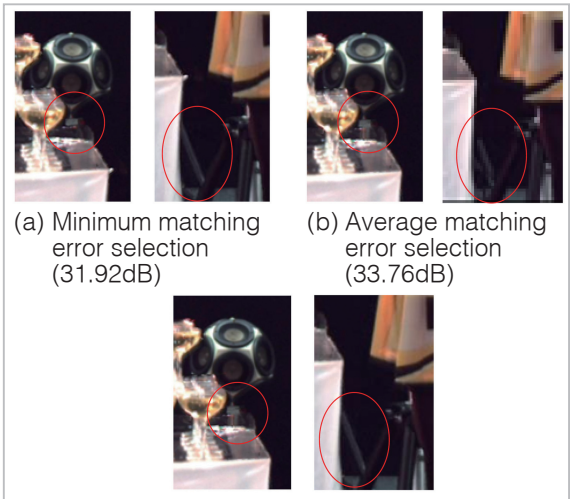


Fig.13 *View 40 synthesized from View 39, 41*

Fig.14 *Difference between synthesized view and camera view*

value selection for the matching error.

In the minimum error selection, the depth value of the edge part of the table has been sharply derived, which confirms the resistance to occlusion. Reversely, the depth value of the glass is rather coarse which seems to be affected by the pseudo-matching.

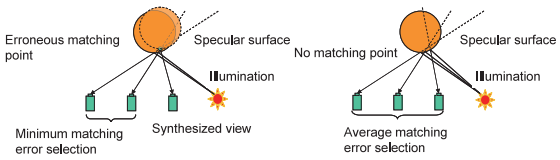
In the average error selection, the depth value of the glass is finely derived, which confirms the resistance to the pseudo-matching. Adversely, the blurring is observed at the depth value of the edge part of the table, which indicates vulnerability to occlusion. The large difference in the depth values of the background between on the left and right sides results from the strong influence of the depth value of the neighboring parts with texture in the smoothing process for determining the depth value due to the plainly black background with little texture that does not lead to a large matching error for any depth value. In regard to parts without texture, even if the intermediate view is synthesized by erroneous depth value, the degradation of image quality can be hardly detected. However, if the depth value of dark objects such as the speaker pole in the scene is erroneous, there is the problem of the disappearance of the background and the appearance of the double image.

When adaptively selecting the minimum error or the average error, the depth value of the glass is finely derived, which indicates the resistant to the pseudo-matching. Also, as to the edge part of the table, in the parts where there is a texture such as the speaker pole, the expanse of the depth value of the table cannot be observed, which confirms the resistance to occlusion.

The depth maps of View 41 and View 39 that were similarly obtained, and the results of synthesizing the center camera image (View 40) from these camera images and the depth mapsare shown in Fig. 13. For the view synthesis, the reference software (VSRS3.5) of MPEG/3DV was used. This software creates the depth map of the intermediate view by projection from the depth maps of the let and right views and their parameters and projects the pixels corresponding to the depth value from the left and right views to synthesize the intermediate view.

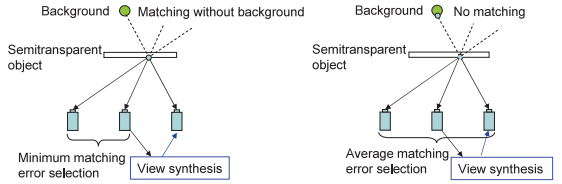
The peak signal to noise ratio (PSNR) of the view synthesized from each of (a) the minimum error selection, (b) the average error selection and (c) the adaptive error selection is each 31.92 dB, 33.76 dB and 32.65 dB, and the PSNR of the synthesized image from the depth map in the average error selection was the highest. This SN difference comes mainly from the difference in accuracy of the depth value of minute parts such as glasses. While the subjective picture quality of the image synthesized by the minimum error selection is high, the errors are large, as shown in Fig. 14.

As shown in Fig. 15(a), this error stems from that the position of illuminating ray specularly-reflected on the glass differs per glass. In the minimum matching error selection, as Fig.



(a) Minimum matching error selection (b) Average matching error selection

Fig.15 *Matching for specular reflection*



(a) Minimum error selection (b) Average error selection

Fig.16 *Depth estimation of semitransparent object*

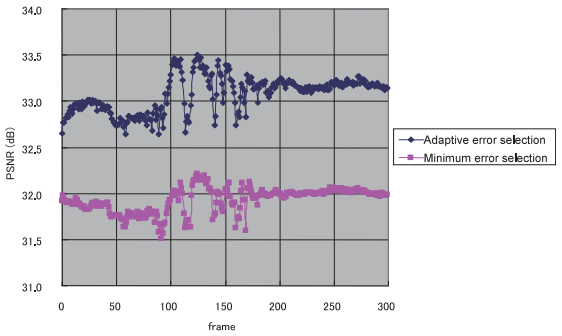


Fig.17 *PSNR (dB) of the synthesized image of Champagne Tower 300 frames*

15(a) shows, since the depth value is estimated in the way that a different right spot position matches per camera, it does not become correct corresponding point matching and the estimated depth value is erroneous. Thus, the image synthesized by using this slight deviates from the camera-captured image, which leads to deteriorated SN.

In the average matching error selection, there is no depth value at which the bright spots of the 3 camera images simultaneously match and any estimated depth value has similar matching errors. When smoothing this matching error by Graph Cuts, the depth value is determined in the way that the difference from the depth value of the neighboring pixels that has been correctly matched becomes small, the depth value of bright spot also becomes almost correct.

Conversely, in the image synthesized from the depth map by the average error selection, although it is observed that the leg of the rightmost speaker at the table becomes a double image, the PSNR is not hugely affected due to the dark object. The double image was caused because the matching error did not grow high even at the incorrect depth value due to the texture with the dark leg and the depth value of the neighboring table had a profound effect in the smoothing process of the depth value to lead to the erroneous depth value of the leg part of the speaker.

Furthermore, the right-hand speaker pole viewed through the semitransparent film wrapped around the table is halfway to disappearing in the minimum error selection, but neither in the average error selection nor in the adaptive error selection. This is, as shown in Fig. 16, because regarding the depth value of the part at which the speaker pole and the semitransparent film overlap with each other in the synthesized view, while the depth value of the semitransparent film is prioritized in the minimum error selection, the depth value of the semitransparent film was not properly obtained in the average error selection and converged at a depth value at which the background poles do not overlap in the 3 camera images.

In the view synthesized from the depth map of the adaptive error selection, although the PSNR slightly decreased, the speaker pole behind the semitransparent film is properly synthesized and the problem of the double image of the speaker pole leg to the right of the table is also resolved.

Figure 17 gives a graphic representation of the PSNR change of View 40 that is synthesized after acquiring the depth of View 39, 41 from the Champagne Tower sequence 300 frames in the adaptive error selection. The adaptive depth estimation method studied at

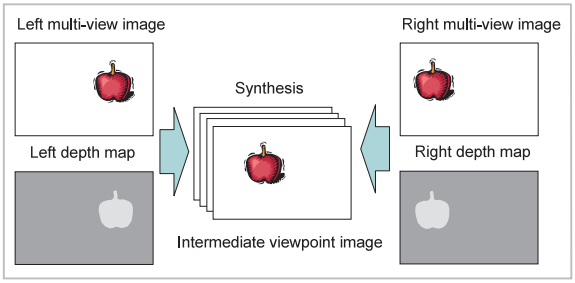


Fig.18 *Multi-view image synthesis form depth images*

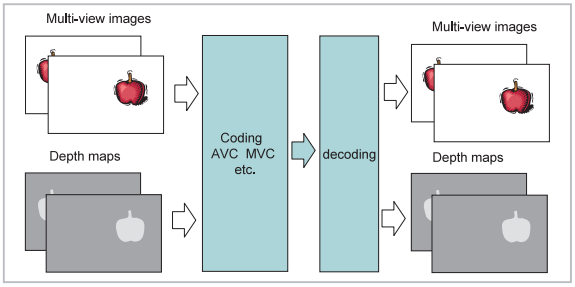


Fig.19 *MVD base coding*

NICT provides the PSNR about 1 dB higher than the minimum error selection method, and thus its efficacy is confirmed.

3.3 Coding

In the following, we will discuss how to code and decode the estimated depth map and the viewpoint image and how to synthesize the intermediate viewpoint image from the decoded viewpoint image and the depth map, and what is considered in MPEG.

3.3.1 Coding of multi-view images and depth map

When the depth map of multi-view image is obtained, as shown in Fig. 18, the intermediate viewpoint image can be synthesized by projective transformation, all the viewpoint images do not need to be coded, it is enough to code only a small number of viewpoint images and their depth maps. If the accuracy of the depth map is high enough, the quality of the synthesized viewpoint image is also high. However, if the accuracy of the depth map acquired on the decoding side is not high enough, the error of the synthesized viewpoint image also needs to be coded.

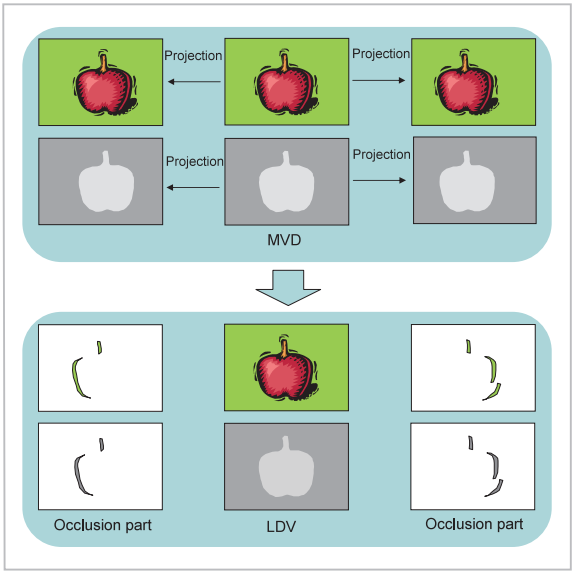


Fig.20 *Conversion from MVD to LDV*

For this coding of viewpoint images and the depth map, as Fig. 19 shows, the existing image coding (AVC, MVC etc.) can be utilized. When using AVC, each viewpoint image and each depth map are coded as individual video image streams. When using MVC, since coding proceeds with the disparity-compensated prediction being conducted between viewpoint images or between depth maps, the coding efficiency is improved, however the viewpoint image and the depth map are separately coded. These are collectively called the multi-view image and depth base coding (MVD: Multi View Depth).

3.3.2 Coding of layered depth image

Just as the intermediate viewpoint image can be projected using the depth map, the viewpoint image and the depth map to be coded itself, as shown in Fig. 20, can be also projected from one viewpoint image or one depth map. Only the occlusion part cannot be projected. Thus, if only the occlusion part is additionally coded, the data to be coded can be largely reduced. This is called the layered depth image base coding (LDV: Layered Depth Video). In regard to coding of LDV data, each viewpoint image, each depth map and the data of each occlusion part can be individually coded by AVC or MVC.

Although this method is valid for the case

where the disparity between viewpoint images is not so large such as the image for multi-view display, the coding efficiency is down in the cases of the large baseline length and the large disparity. In either case, the bit rate required for the depth map is about one fifth of the viewpoint image. Therefore, when using the depth map, there is little difference in the coding efficiency from the existing methods (MVC etc.) with respect to coding of stereoscopic images. However, since the code size does not increase as far as a certain number of viewpoints, there is the characteristic that the more viewpoints, the more the coding efficiency improves.

3.4 Viewpoint image synthesis

In regard to the synthesizing of the intermediate viewpoint images on the decoding side, when directly projecting viewpoint images on the both sides using their depth maps, not all the pixels of the synthesized image are filled and the unprojected pixels leave holes. When these are filled by the median filter or the impainting method that copies the neighboring pixels, the distortion may occasionally become higher in the object with partially different colors and textures. As an alternative, as Fig. 21 shows, create the depth map of the synthesized image at the viewpoint position by the projection from the depth maps on the both sides, and project the pixels corresponding to that depth from the viewpoint images on the both sides after likewise filling the holes of this depth map. Then, since the depth of the same object does not drastically vary, the synthesized image has little distortion. In order to further reduce the noise of the synthesized image, the object boundary is filtered and the inter-frame filter is applied for still parts.

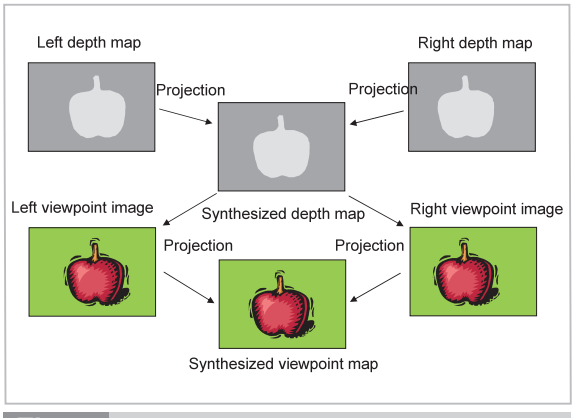


Fig.21 *Synthesis of viewpoint images by synthesized depth map*

4 Conclusion

In the above, the coding method of multi-view image that the MPEG group affiliated with ISO has been standardizing was introduced, and the adaptive depth estimation investigated at NICT discussed. The multi-view image is the fundamental element for 3D images and free viewpoint images, which would seem to become more important in the days to come.

MVC that takes only the multi-view image as the input and conducts the disparity-compensated prediction between viewpoint images prediction by translation of pixel block has already been standardized and begun to be put into practical use. 3DV/FTV that aims to extract the disparity per pixel from the multi-view image as the depth map and to achieve a higher coding efficiency seems to be standardized and used for glasses-free 3D images and free viewpoint images from hereon in. Among these, the depth map is not only valid for improving the coding efficiency of the multi-view image, but also an important element for generating 3D images such as the electronic holography and free viewpoint images. For successful applications of these, it is essential to enable us to easily obtain the high-quality depth map, which has been studied all over the world. We hope to expect future development.

Acknowledgements

The multi-view images “Akko & Kayo” and “Champagne Tower” utilized in this report were captured in the Tanimoto Research Laboratory at Nagoya University. Yasuda and Aoki Research Laboratory at the University of Tokyo kindly provided help to the capture of “Akko & Kayo”. The authors would like to express a deep sense of gratitude to all the parties concerned.

References

1 K. Enami, M. Okui, and N. Inoue, "NICT's Research Strategy on Ultra Realistic Communication," Proc. IIEEJ, 06-03 (Proc. ITE, Vol. 30, No. 58/Proc. IEICE Vol.106, No. 338/Proc. IEE, EDD-06-75-85), pp. 1-6, 2006.

2 T. Honda, "Final Report on Advanced Three-Dimensional Video Communication Project," TAO, 1997.

3 M. Tanimoto, "Overview of Free View-point Television," Signal Processing: Image Communication, Vol. 21, No. 6, pp. 454-461, 2006.

4 ISO/IEC 14496-10, or ITU-T H.264, 2009.

5 M. Tanimoto, "International Standard Technology Research on Free Viewpoint Television Transmission Method," Strategic Information and Communications R&D Promotion Programme, Proc. No. 5 Reporting Conference, 2008.

6 N. Fukushima, K. Matsumoto, T. Endoh, T. Fujii, and M. Tanimoto, "Rectification Method for Two-dimensional camera Array by Using Parallelizing Locus of Feature Points," Journal ITE, Vol. 62, No. 4, pp. 564-571, 2008.

7 T. Senoh, K. Yamamoto, R. Oi, T. Mishina, and M. Okui, "Computer Generated Electronic Holography of Natural Scene from 2D Multi-view Images and Depth Map," Proc. of 2nd International Symposium on Universal Communication (ISUC) 2008, pp. 126-133, 2008.

8 K. Palaniappan, et.al: Robust Stereo Analysis, Computer Vision, Proc. International Symposium on Digital Object Identifier, pp. 175-181, 1995.

9 J. Sun, N. Zheng, and H. Shun, "Stereo matching using belief propagation," IE^3 Trans. Pattern Analysis and Machine Intelligence, Vol. 25, No. 7, pp. 787-800, 2003.

10 Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," IE^3 Trans. Pattern Analysis and Machine Intelligence, Vol. 23, No. 11, pp. 1222-1239, 2001.

11 T. Naemura and H. Harashima, "Occlusion-Free Estimation of Disparity from Multi-View Images," Proc. IEICE, SD-7-1, 1996.

12 H. Imaizymi, M. Katayama, and Y Iwadate, "Depth estimation Algorithm for Multi-ocular Images and representation of Intermediate Viewpoint," Proc. ITE, IE2001-59, PRMU2001-79, MVE2001-58, pp. 109-116, 2001.

13 http://www.tanimoto.nuee.nagoya-u.ac.jp/

14 ISO/IEC JTC1/SC29/WG11, "Draft Report on Experimental Framework for 3D Video Coding," N11273, 2010.

(Accepted Sept. 9, 2010)

SENOH Takanori, *Dr. Eng.*

Expert Researcher, 3D Spatial Image and Sound Group, Universal Media Research Center

Electronic Holography, 3D Image Technology

YAMAMOTO Kenji, *Dr. Eng.*

Senior Researcher, 3D Spatial Image and Sound Group, Universal Media Research Center

Electronic Holography, 3D Image Technology

OI Ryutaro, *Dr. Sci.*

Senior Researcher, 3D Spatial Image and Sound Group, Universal Media Research Center

Optical Wave Propagation Analysis, Holography, 3D Imaging Technology, Image Sensor

KURITA Taiichiro, *Dr. Eng.*

Group Leader, 3D Spatial Image and Sound Group, Universal Media Research Center

Television System, Information Display, 3D Image Technology