INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11
MPEG2005/M12485
October 2005, Nice

Title: Generation and Coding of Layered Depth Images for Multi-view Video

Source: GIST and ETRI
Authors: Yo-Sung Ho, Seung-Uk Yoon, Eun-Kyung Lee, and Sung-Yeol Kim
(Gwangju Institute of Science and Technology)
Kugjin Yun, Daehee Kim, Namho Hur, and Soo-In Lee
(Electronics and Telecommunications Research Institute)

Status: Proposal

1 Introduction

Layered depth image (LDI) is an efficient approach to represent three-dimensional (3-D) objects with complex geometry for image-based rendering (IBR). LDI has already presented as a useful tool for multi-texturing and IBR in MPEG-4 AFX CE A8 [1]. In AFX, the functionality of LDI is mainly focused on texturing and rendering with depth. However, the objective of IBR is helpful to the multi-view video coding. In this document, we describe the generation and coding of LDI for multi-view video with depth information as an effort for MPEG-4 3DAV.

2 Analysis of the MPEG 3DAV Multi-view Video Test Data

There are various kinds of multi-view video test sequences provided by MPEG AHG on 3DAV [2][3]. Several proponents provide about 20 sequences with different properties. Recently, MPEG has been issued the call for proposals (CfP) on MVC [4] and consequently eight sequences have been selected as the test sets for CfP on MVC. They have been selected by considering the variety of features, such as the number of cameras (5, 8, and 100), camera arrangements (1-D parallel, 1-D parallel convergent, 1-D arc, 2-D cross, and 2-D array), frames per second (15, 25, and 30), image resolutions (VGA and XVGA), scene complexity, and camera motions. Besides, all the test sequences contain camera parameters for their own camera arrangement and the validation of those camera parameters has been performed in MPEG AHG on 3DAV. The properties of the multi-view video test sequences are summarized in Table 1 and Table 2 [4]. Among them, Microsoft Research (MSR) provided the multi-view video sequence, Breakdancers with camera parameters and depth information [5][6].

Table 1. Properties of MPEG 3DAV test sequences (A: available, N/A: not available)

Property	KDDI	MERL	HHI	Nagoya	MSR
Sequences	Flamenco Objects Crowd Golf, Race	Ballroom Exit	Jungle Uli	Rena Akiko Akko&Kayo	Breakdancers Ballet
Nunmber of Cameras	5/8	8	8	100	8
Camera Parameters	A	A	A	A	A
Depth Information	N/A	N/A	N/A	N/A	A

Table 2. Test Data Sets for CfP on MVC

Data Set	Sequences	Imag Property	Number of Camera	Camera Arrangement
MERL	Ballroom	VGA, 25fps	8	1-D parallel
	Exit	VGA, 25fps	8	1-D parallel
KDDI	Race1	VGA, 30fps	8	1-D parallel
	Flamenco2	VGA, 30fps	5	2-D parallel (cross)
HHI	Uli	XVGA, 25fps	8	1-D parallel convergent
MSR	Breakdancers	XVGA, 15fps	8	1-D arc
Nagoya Univ.	Rena	VGA, 30fps	100	1-D parallel
	Akko&Kayo	VGA, 30fps	100	2-D array

There are several problems in generating LDI frames from the test sets without depth information. Although we can easily compute depth images from disparity maps under the parallel camera arrangement, the quality of the computed depth map is not sufficient. If we use a more stable and accurate method and perform more iterations or refinements, we can obtain more reliable results. However, it is very time consuming and it requires another preprocessing and postprocessing to get sufficient quality of depth maps.

Because of these reasons, we are now focusing on the test set from MSR [5][6]. MSR data include a sequence of 100 images captured from eight cameras; the camera arrangement is 1-D arc with about 20cm horizontal spacing. Depth maps computed by stereo matching algorithms are provided for each camera together with the camera parameters: intrinsic parameters, barrel distortion, and rotation matrix. The exact depth range is also included.

3 Generation of Layered Depth Images from Multi-view Video with Depth Information

As we described in “Multi-view Video Coding using Layered Depth Image” [7], LDI can be generated by warping multiple depth images into an LDI view. However, there were several problems to warp those depth images, because the warping equation is only useful for the 3-D graphics models. In this experiment, we use the test sequences with depth information (MSR data) provided by MPEG 3DAV. We have adopted the following incremental 3-D warping equation [8],

$$C_1 = V_1 \cdot P_1 \cdot A_1,\ C_2 = V_2 \cdot P_2 \cdot A_2,\ T_{1,2} = C_2 \cdot C_1^{-1} \quad (1)$$

$$T_{1,2} \cdot \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ 1 \end{bmatrix} = \begin{bmatrix} x_2 \cdot w_2 \\ y_2 \cdot w_2 \\ z_2 \cdot w_2 \\ w_2 \end{bmatrix} = T_{1,2} \cdot \begin{bmatrix} x_1 \\ y_1 \\ 0 \\ 1 \end{bmatrix} + z_1 \cdot T_{1,2} \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = start + z_1 \cdot depth \tag{2}$$

where V is the viewport matrix, P is the projection matrix, and A is the affine matrix.

Each matrix of the incremental warping equation can be easily found in scenes generated by computer graphics. The viewport matrix is computed from the image resolution; the projection matrix is automatically determined by OpenGL according to the orthogonal or perspective view; and the affine matrix is computed by the rotation and translation matrix. However, it is difficult to calculate these matrices in the natural video because the meanings of P and A are not defined clearly [9].

Because of these reasons, we use a different camera matrix calculated from the given camera parameters of the MPEG-4 3DAV test sequences, instead of estimating each V, P, and A matrix in the natural video. The camera matrix is as follows.

$$\dot{C}_1 = \dot{A}_1 \cdot \dot{E}_1,\ \dot{C}_2 = \dot{A}_2 \cdot \dot{E}_2,\ \dot{T}_{1,2} = \dot{C}_2 \cdot \dot{C}_1^{-1} \tag{3}$$

$$\dot{A} = \begin{bmatrix} -f_{S_x} & \theta & t_x \\ 0 & -f_{S_y} & t_y \\ 0 & 0 & 1 \end{bmatrix},\ \dot{E} = \begin{bmatrix} R_{11} & R_{12} & R_{13} & T_1 \\ R_{21} & R_{22} & R_{23} & T_2 \\ R_{31} & R_{32} & R_{33} & T_3 \end{bmatrix} \tag{4}$$

$$\dot{T}_{1,2} \cdot \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ 1 \end{bmatrix} = \begin{bmatrix} x_2 \cdot w_2 \\ y_2 \cdot w_2 \\ z_2 \cdot w_2 \\ w_2 \end{bmatrix} = \dot{T}_{1,2} \cdot \begin{bmatrix} x_1 \\ y_1 \\ 0 \\ 1 \end{bmatrix} + z_1 \cdot \dot{T}_{1,2} \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = start + z_1 \cdot depth \tag{5}$$

where f_{S_x}, f_{S_y} are the focal length, S_x, S_y are scaling factors, t_x, t_y are positions of the focal center, and θ is the skew angle. $\dot{A}$ defines the intrinsic camera parameters and $\dot{E}$ is the Euclidean transform expressing a rotation and a translation. Finally, The camera matrix C is changed to $\dot{C}$, but the Eq. 5 is the same as Eq. 1. We should add an additional row $[0 \quad 0 \quad 0 \quad 1]$ to make a homogeneous 4 x 4 camera matrix $\dot{C}$, because $\dot{A} \cdot \dot{E}$ becomes a 3 x 4 matrix. We perform the incremental 3-D warping using the above modified camera matrix and the warping result is depicted in Fig. 1.

(a) From Cam. 0 to Cam. 4 (b) From Cam. 1 to Cam. 4 (c) From Cam. 7 to Cam. 4

Fig. 1. Results of the incremental 3-D warping

We can observe that actors are slightly rotating as the camera number changes. In order to identify the warping results clearly, we do not interpolate holes. In Fig. 1, camera

number 4 is the reference LDI view and the warping was performed from other camera locations to the reference LDI view. In detail, there are holes in the right portion of the actors when we move from the left cameras of the reference LDI view. On the other hand, holes occur in the left portion of the actors when we move from the right cameras of the reference LDI view.

4 Encoding of the Constructed LDI Frames

Before encoding the generated LDI frames, we have analyzed the properties of the constructed LDI frames [10]. LDI pixel contains color values, depth between the camera and the pixel, and other attributes that support rendering of LDI. Three key characteristics of LDI are: (1) it contains multiple layers at each pixel location, (2) the distribution of pixels in the back layer is sparse, and (3) each pixel has multiple attribute values. Because of these special features, LDI enables us to render arbitrary views of the scene at new camera positions.

Because each layer of the constructed LDI has different number of pixels, we need to aggregate scattered pixels into the horizontal or vertical direction [10]. Although H.264/AVC is powerful to encode rectangular images, it does not support shape-adaptive encoding modes. We therefore adapt each layer to fit H.264/AVC by using data aggregation and reordering of the aggregated images. First, the scattered pixels in each layer are pushed to the horizontal direction. Second, the images containing collected pixels are merged into a single image. Finally, the generated one is reordered and divided into the images with pre-defined resolutions to employ H.264/AVC. Figure 2 represents the results of the data aggregation with horizontal directions.

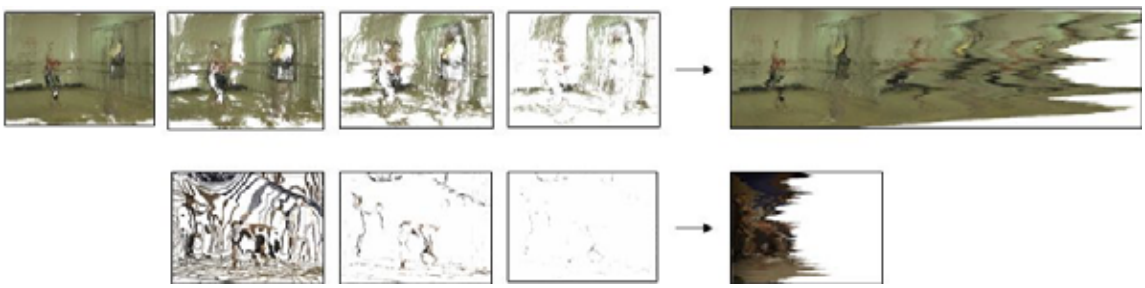

Fig. 2. Results of the data aggregation with horizontal direction

5 Experimental Results and Analysis

After MPEG-4 3DAV test sequences have been released, there have been lots of discussions about imperfection of the test materials, and more robust and accurate test data sets are solicited. Although we are currently working on those test sequences, we expect that more accurate test sequences with additional information could be available soon. In our experiment, we have used the test sequences from Microsoft Research with the incremental 3-D warping equation to generate LDI.

We have obtained LDI frames from the Ballet and Breakdancers sequences of the MSR data set by 3-D warping with the given depth images. Using eight color and eight depth images of each sequence, we perform incremental warping to construct a single LDI frame. In other words, the first eight color and depth frames of Ballet or Breakdancers sequence for camera zero are used to generate the first LDI frame; the second 16 images are used to

make the second LDI frame; and so on. After that, those LDI frames are processed in our proposed MVC framework [11].

Figure 3 shows the characteristics of each layer of the constructed LDI. For the first LDI frame, there are no holes in the first layer. However, holes are increased as the number of layers increase.

(a) 1st layer (b) 2nd layer (c) 3rd layer (d) 4th layer

Fig. 3. Characteristics of the each layer of the generated LDI frame

In Table 3 and Table 4, we have compared the data size between sum of frames of the test sequence and the generated LDI frame. In each table, sum of frames means that the summation of eight color and depth images of the test sequence. For the Ballet sequence, the encoded LDI frame has more data than sum of frames using the simulcast method. On the other hand, the encoded LDI frame generated from the Breakdancers sequence has fewer amounts of data. The major reason is the depth variation of each sequence; the Ballet sequence has smaller depth variation than the Breakdancers sequence. The more detailed analysis on the relationship between the total bit-rate and the depth variation of the test sequence would be needed in the future.

In addition, the data size has been reduced when the validated camera parameters are used. The registration of each layer was needed before the validation, but we have found that the registration is not necessary when the corrected camera parameters are used. We regenerated LDI frames using the newly validated camera parameters and observed that the data size was decreased as shown in Table 3 and Table 4.

Table 3. Comparison of data size for the Ballet sequence

	1st 8 Frames	2nd 8 Frames
Sum of frames (color + depth) [Kbytes]	25,165.9	25,165.9
LDI frame generated from 16 images [Kbytes]	14,078.0	14,061.6
Simulcast using H.264 (color + depth) [Kbytes]	134.4	149.5
Encoded LDI frame using the data aggregation (using invalidated camera parameters) [Kbytes]	159.7	168.7
Encoded LDI frame using the data aggregation (using validated camera parameters) [Kbytes]	133.7	148.9

Table 4. Comparison of data size for the Breakdanders sequence

	1st 8 Frames	2nd 8 Frames
Sum of frames (color + depth) [Kbytes]	25,165.9	25,165.9
LDI frame generated from 16 images [Kbytes]	12,726.6	12,689.7
Simulcast using H.264 (color + depth) [Kbytes]	165.9	160.6
Encoded LDI frame using the data aggregation (using invalidated camera parameters) [Kbytes]	155.3	151.9
Encoded LDI frame using the data aggregation (using validated camera parameters) [Kbytes]	88.5	85.6

6 Conclusion

In this document, we have described the generation procedure for LDIs from multi-view video with depth information and explained our encoding methods of multi-view video. Through the experiments, we have observed that the LDI framework has a possibility for efficient encoding of multi-view video data. For the next meeting, we will response to the CfP on MVC with sufficient experiments.

7 References

[1] ISO/IEC JTC1/SC29/WG11 N4220, "Animation Framework eXtension Core Experiments Description," July 2001.

[2] ISO/IEC JTC1/SC29/WG11 N6720, "Call for Evidence on Multi-view Video Coding," October 2004.

[3] ISO/IEC JTC1/SC29/WG11 N7094, "Preliminary Call for Proposals on Multi-view Video Coding," April 2005.

[4] ISO/IEC JTC1/SC29/WG11 N7327, "Call for Proposals on Multi-view Video Coding," July 2005.

[5] Interactive Visual Media Group at Microsoft Research, http://www.research.microsoft.com/vision/ImageBasedRealities/3DVideoDownload/

[6] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality Video View Interpolation using a Layered Representation," ACM SIGGRAPH, pp. 600-608, Aug. 2004.

[7] ISO/IEC JTC1/SC29/WG11 m11278, "Multi-view Video Coding using Layered Depth Image," October 2004.

[8] J. Shade, S. Gortler, L. He, and R. Szeliski, "Layered Depth Images," ACM SIGGRAPH, pp. 231-242, July 1998.

[9] ISO/IEC JTC1/SC29/WG11 m11916, "Preliminary Results for Multi-view Video Coding using Layered Depth Image," April 2005.

[10] ISO/IEC JTC1/SC29/WG11 m12278, "Intermediate Result on Multi-view Video Coding using Layered Depth Images," July 2005.

[11] ISO/IEC JTC1/SC29/WG11 m11582, "A Framework for Multi-view Video Coding using Layered Depth Image," January 2005.