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/M11916
April 2005, Busan

Title: Preliminary Results for Multi-view Video Coding using Layered Depth Image
Source: GIST
Authors: Yo-Sung Ho, Seung-Uk Yoon, Sung-Yeol Kim, and Eun-Kyung Lee
(Gwangju Institute of Science and Technology)
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. In this document, we describe preliminary results for multi-view video coding using layered depth image as an effort for MPEG-4 3DAV.

2 Conversion between Multi-view Video and Layered Depth Image

As we described in "Multi-view Video Coding using Layered Depth Image" [2], LDI can be generated by warping multiple depth images into an LDI view. The first step of this conversion is therefore to obtain depth images or depth maps from a multi-view video. Since there are various kinds of multi-view video sequences, it is necessary to analyze the properties of those data sets. After acquiring multiple depth images from various viewpoints, we need to obtain camera parameters at each view position to perform 3-D warping from the depth image to LDI.

There are two types of test video sources provided by MPEG-4 3DAV: dense and sparse test sequences [3]. Among them, KDDI provides 6 dense sequences with two different types of parallel camera configurations. ETH provides sparse 3-D video data with color, shape, and depth [4]. Recently, Microsoft Research (MSR) provided two multi-view video test sequences for 3DAV with some camera parameters [5][6]. In this document, we will focus on the data sets from MSR. The characteristics of the various types of test sequences are summarized in Table 1.

Table 1. Characteristics of MPEG-4 3DAV test sequences: A: available, N/A: not available

Property	KDDI	Aquarium	X'mas	ETH-Zurich	STMicroelectronics	MSR
Camera Configuration	Parallel	1-D arc	101 cameras	Convergent	Convergent	1-D arc
Intrinsic Camera Parameters	A	A	N/A	A	N/A	A
Center of Projection, C	A	N/A	N/A	A	A	A
Projection Matrix, P	A	N/A	N/A	A	A	A
Depth	N/A	N/A	N/A	A	N/A	A
Disparity	N/A	N/A	N/A	N/A	N/A	N/A

MSR data includes a sequence of 100 images captured from eight cameras. Depth maps computed from stereo matching are also included for each camera along with the calibration parameters: intrinsic parameters, barrel distortion, and rotation matrix. In addition, the depth range is provided. We depicted some of sample depth images from the MSR data set in Fig. 1. As we can see in Fig. 1, depth images from MSR are more reliable than those of KDDI [7], but they still have some artifacts and blurred portions in each depth image.

(a) (b)

Fig. 1. Depth images from MSR (1024x768, 8 bpp): (a) Ballet, (b) Breakdancers

After analyzing the properties of those test data sets, we perform 3-D warping to generate a single LDI using multiple depth images. We use the following incremental 3-D warping equation, mentioned by Shade [8]

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

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

where V is the viewport matrix, P is the projection matrix, and A is the affine matrix. In Fig. 2, we shows the camera configuration. We can generate a new view at $\dot{C}_2$ by 3-D warping from the camera location $\dot{C}_1$. Finally, we can obtain the LDI frame by depth comparison and thresholding [8].

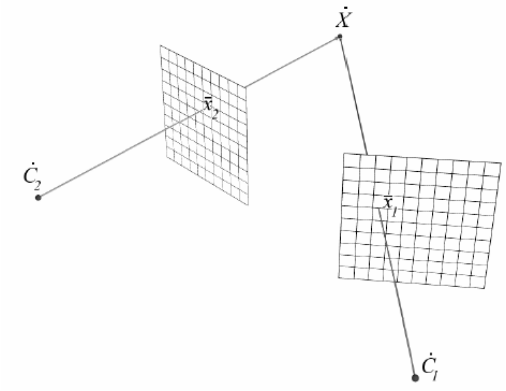

Fig. 2. Camera configuration of 3-D warping

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. We therefore use the given camera matrix of the MPEG-4 3DAV test sequences instead of estimating each V, P, and A matrix in the natural video.

As we already mentioned in "Framework for Multi-view Video Coding using Layered Depth Image" [7], if we use the KDDI data sets as the natural video input, we need to estimate the center of projection of a camera and the projection matrix P. Basis vectors are then calculated using the estimated position of the camera and the matrix P. Disparity can be computed by a stereo matching algorithm, and depth values are directly obtained from disparity in the parallel camera configuration.

On the other hand, if we use the ETH-Zurich test sequence, we do not have to estimate the center of projection and the projection matrix P; however, we need to compute the disparity δ in the McMillan's equation [9]. It is not a simple task to obtain the disparity from the given depth because 16 cameras are not parallel, but convergent. The well-known formula, d = bf/z, where d is disparity, b is baseline, f is focal length, and z is depth, cannot be directly applied to calculate the disparity from the depth in the non-parallel camera configuration.

Because of these reasons, we are now focusing on MSR test data sequences with the incremental warping equation. In the original McMillan's warping equation, there are many parameters to be computed and the calculation of δ from depth is a difficult problem. However, if we use the incremental warping equation, we can avoid those problems and reduce the computational complexity for 3-D warping, which is well described in the Shade's paper [8].

3 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 MSR test data with the incremental 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 [7][10].

In Table 2 and Table 3, 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. Because MSR data contains 100 color and depth images in the format of "bmp" for each camera, we have converted the format of each image to "yuv 4:2:0" format, which is appropriate for H.264.

Table 2. 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

Table 3. 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

4 Conclusion

In this document, we have described the characteristics of the 3DAV test sequences and preliminary results for multi-view video coding using layered depth image. We can encode multi-view video sequences with the depth information efficiently using the concept of LDI. As we described in this document, we have generated LDI frames from natural video sequences and have seen that our LDI framework has a possibility for efficient coding of multi-view video data. For the next meeting, we will present more sufficient experimental results based on our LDI framework.

5 References

[1] ISO/IEC JTC 1/SC 29/WG 11/N4220, "Animation Framework eXtension Core Experiments Description," July 2001.

[2] ISO/IEC JTC 1/SC 29/WG 11/M11278, "Multi-view Video Coding using Layered Depth Image," Oct. 2004.

[3] ISO/IEC JTC 1/SC 29/WG 11/N6720, "Call for Evidence on Multi-view Video Coding," Oct. 2004.

[4] 3-D Video at ETH Zurich, http://graphics.ethz.ch/3dvideo/main.php

[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 JTC 1/SC 29/WG 11/M11582, "A Framework for Multi-view Video Coding using Layered Depth Image," Jan. 2005.

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

[9] L. McMillan, "An Image-based Approach to Three-Dimensional Computer Graphics," Ph.D. Dissertation, 1997.

[10] ISO/IEC JTC 1/SC 29/WG 11/M11279, "Coding of Layered Depth Image using Coherency between Point Samples," Oct. 2004.