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/M12278
July 2005, Poznan

Title: Intermediate Result on Multi-view Video Coding using Layered Depth Images
Source: GIST and ETRI

Authors: Yo-Sung Ho, Seung-Uk Yoon, Eun-Kyung Lee, Sung-Yeol Kim, and Seung-Hwan 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. In this document, we describe intermediate result on multi-view video coding using layered depth images as an effort for MPEG-4 3DAV.

2 Conversion between Multi-view Video and Layered Depth Images

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 the 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 many kinds of multi-view video test sequences provided by MPEG AHG on 3DAV [3][4]. Among them, KDDI provides 10 dense sequences using two different types of parallel camera arrangements. Recently, Microsoft Research (MSR) provided two multi-view video sequences for 3DAV with camera parameters and depth information [5][6]. The properties of the various types of test sequences are summarized in Table 1.

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

Property	KDDI	MERL	HHI	Nagoya	MSR
# of Cameras	5/8	8	8	100	8
Camera Configuration	Parallel	1-D Parallel	1-D Parallel	1-D arc/parallel	1-D arc
Camera Parameters	A	A	A	A	A
Depth Information	N/A	N/A	N/A	N/A	A

MSR data include a sequence of 100 images captured from eight cameras. Depth maps computed from stereo matching methods are also included for each camera together with the calibration parameters: intrinsic parameters, barrel distortion, and rotation matrix. In addition, the depth range is provided.

After analyzing the properties of those test data sets, we perform 3-D warping to generate LDI using multiple depth images. We use the incremental 3-D warping equation [7]. Each matrix of the equation can be easily found in scenes generated by computer graphics. However, it is difficult to calculate these matrices in the natural video because the meanings of some matrices are not defined clearly. We therefore use the given camera matrix of the MSR data sets instead of estimating each matrix in the natural video [8].

As we already mentioned before, if we use the KDDI data sets as the natural video input, we need to estimate disparity and depth values. Although we can easily compute depth images from disparity maps using the parallel configuration of cameras, the quality of the computed depth map is not sufficient [9].

Because of these reasons, we are now focusing on MSR test data sequences with the incremental 3-D warping equation. In the original McMillan's warping equation [10], there are many parameters to be computed and the calculation of disparity 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 [7].

3 Characteristics of the Generated LDI Frames

Before encoding the generated LDI frames, we first analyze the properties of the constructed LDI frames. 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. Table 2 lists the number of pixels included in each layer of the generated LDI. The LDI frame in Table 2 is constructed from the first 8 frames of the Ballet sequence.

Table 2. Pixel distribution of the generated LDI

Layer	Pixel Occupation [%]	Layer	Pixel Occupation [%]
1	100	5	16.5
2	98.9	6	3.2
3	84.2	7	0.3
4	43.7	8	0.0

4 Encoding of the Constructed LDI Frames

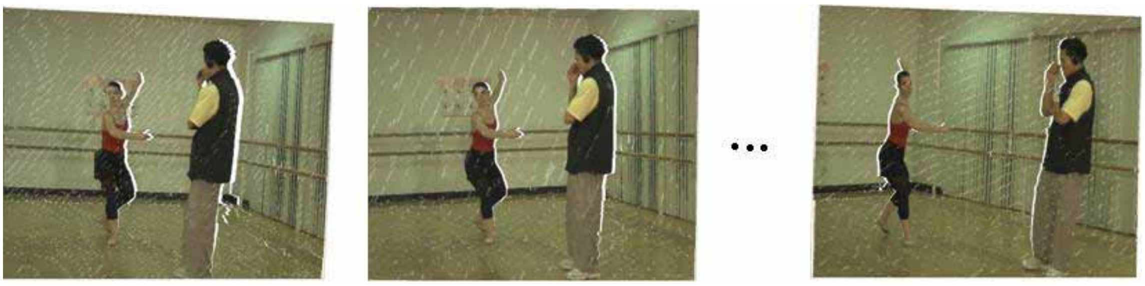
Because each layer of the constructed LDI has different number of pixels as shown in Table 2, we need to aggregate scattered pixels into the horizontal or vertical direction. Although H.264/AVC is powerful to encode rectangular images, it does not support shape-adaptive encoding modes. We therefore adapt 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. At the moment, we have only considered the aggregation with the horizontal axis and a simple reordering method, but we will further explore other kinds of approaches.

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 [8].

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 [8][9].

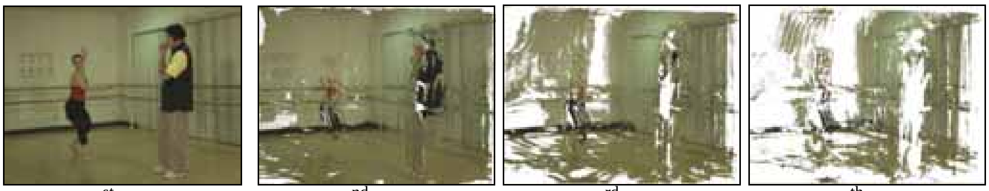
Figure 1 shows the results of 3-D warping using the constructed camera matrix. We can observe that actors are slightly rotating as the camera number changes. In order to identify the warping results clearly, we do not interpolate holes. In Fig. 1, camera number 4 is the reference LDI view and the warping 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.



(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

Figure 2 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

(e) 5th layer (f) 6th layer (g) 7th layer

Fig. 2. 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 bitrate and the depth variation of the test sequence would be needed in the future.

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 proposed methods [Kbytes]	159.7	168.7

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 proposed methods [Kbytes]	155.3	151.9

6 Conclusion

In this document, we have briefly described the characteristics of the newly provided 3DAV test sequences according to “Preliminary Call for Proposals for Multi-view Video Coding” [8] and intermediate result on MVC using layered depth images. Although we have performed several experiments on MVC using our framework in the limited environments, we 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.

7 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] ISO/IEC JTC 1/SC 29/WG 11/N7094, “Preliminary Call for Proposals on Multi-view Video Coding,” April. 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] J. Shade, S. Gortler, L. He, and R. Szeliski, “Layered Depth Images,” ACM SIGGRAPH, pp. 231-242, July 1998.

[8] ISO/IEC JTC 1/SC 29/WG 11/M11916, “Preliminary Results for Multi-view Video Coding using Layered Depth Image,” April 2005.

[9] ISO/IEC JTC 1/SC 29/WG 11/M11582, “A Framework for Multi-view Video Coding using Layered Depth Image,” Jan. 2005.

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