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/M11582
January 2005, Hong Kong

Title: A Framework for Multi-view Video Coding using Layered Depth Image
Source: GIST

Authors: Yo-Sung Ho, Seung-Uk Yoon, and Sung-Yeol Kim
(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 propose a framework to encode multi-view video using the concept of LDI.

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: KDDI-A and KDDI-B. KDDI-A is one-dimensional parallel with 8 cameras and KDDI-B is two-dimensional parallel with 5 cameras. ETH provides sparse 3-D video data with color, shape, and depth [4]. In this document, we will focus on the data sets from KDDI and ETH.

Although camera parameters provided by KDDI contain 20cm baseline, 1/3″ Sony CCD, 4/6/8mm focal length, and M12 microlenses, there is no record about which focal length was applied when shooting. Extrinsic camera parameters did not belong to KDDI data sets. This causes some problems when we convert a multi-view video to LDI because we need to estimate the center of projection C and the projection matrix P for 3-D warping.

On the other hand, the ETH-Zurich data set contains explicit camera center locations, radial distortion coefficients, and a projection matrix P including intrinsic camera parameters. ETH offers video sequences with three types of data for 16 camera locations in the convergent camera configuration. The characteristics of the two types of test sequences are summarized in Table 1.

Table 1. Characteristics of KDDI and ETH-Zurich test sequences

Property	KDDI	ETH-Zurich
Camera Configuration	Parallel	Convergent
Intrinsic Camera Parameters	Yes	Yes
Center of Projection, C	No	Yes
Projection Matrix, P	No	Yes
Depth	No	Yes
Disparity	No	No

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 3-D warping equation proposed by McMillan [5]

$$
\begin{aligned}
u_2 &= \frac{\bar{a}_1 \cdot (\bar{b}_2 \times \bar{c}_2) u_1 + \bar{b}_1 \cdot (\bar{b}_2 \times \bar{c}_2) v_1 + \bar{c}_1 (\bar{b}_2 \times \bar{c}_2) + (\dot{C}_1 - \dot{C}_2) \cdot (\bar{b}_2 \times \bar{c}_2) \delta(u_1, v_1)}{\bar{a}_1 \cdot (\bar{a}_2 \times \bar{b}_2) u_1 + \bar{b}_1 \cdot (\bar{a}_2 \times \bar{b}_2) v_1 + \bar{c}_1 (\bar{a}_2 \times \bar{b}_2) + (\dot{C}_1 - \dot{C}_2) \cdot (\bar{a}_2 \times \bar{b}_2) \delta(u_1, v_1)} \\
v_2 &= \frac{\bar{a}_1 \cdot (\bar{c}_2 \times \bar{a}_2) u_1 + \bar{b}_1 \cdot (\bar{c}_2 \times \bar{a}_2) v_1 + \bar{c}_1 (\bar{c}_2 \times \bar{a}_2) + (\dot{C}_1 - \dot{C}_2) \cdot (\bar{c}_2 \times \bar{a}_2) \delta(u_1, v_1)}{\bar{a}_1 \cdot (\bar{a}_2 \times \bar{b}_2) u_1 + \bar{b}_1 \cdot (\bar{a}_2 \times \bar{b}_2) v_1 + \bar{c}_1 (\bar{a}_2 \times \bar{b}_2) + (\dot{C}_1 - \dot{C}_2) \cdot (\bar{a}_2 \times \bar{b}_2) \delta(u_1, v_1)}
\end{aligned}
\tag{1}
$$

where $\dot{C}_1$, $\dot{C}_2$ are camera positions, ($\bar{a}_1, \bar{b}_1, \bar{c}_1$), ($\bar{a}_2, \bar{b}_2, \bar{c}_2$) are basis vectors, and δ is disparity. (u_1, v_1) is a pixel coordinate of an image plane at $\dot{C}_1$ and (u_2, v_2) is that at $\dot{C}_2$. We can generate a new view at $\dot{C}_2$ by means of 3-D warping from the camera location $\dot{C}_1$ to $\dot{C}_2$. Finally, we can obtain the LDI frame by depth comparison and thresholding [6]. Figure 1 shows the procedure for generating LDI from multi-view video sequences.

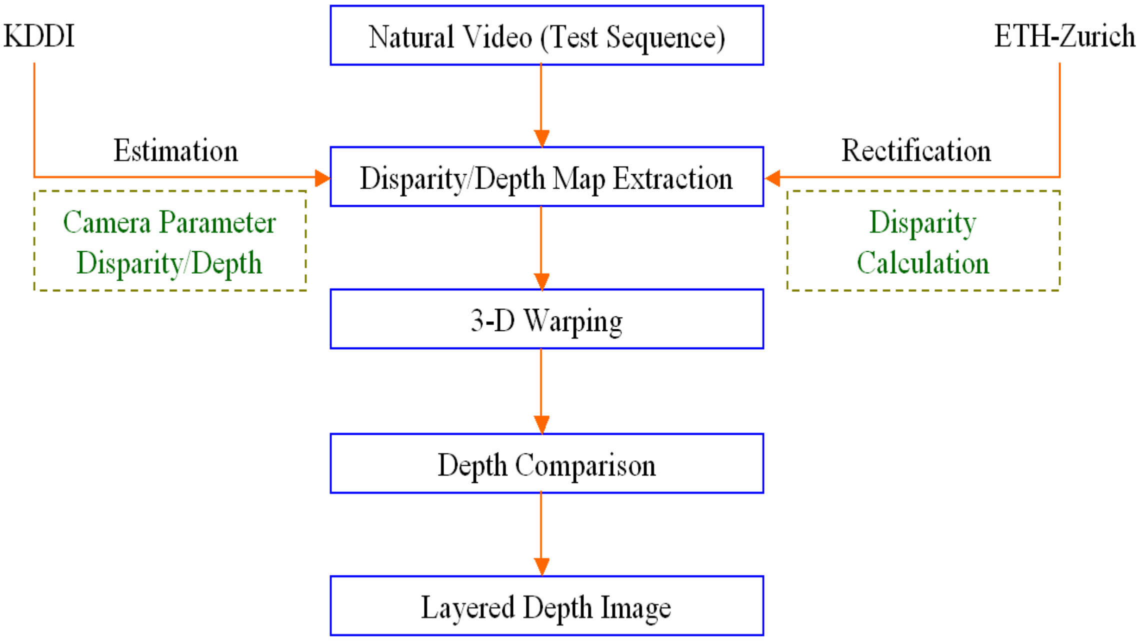


Fig. 1. Generation of LDI from natural multi-view video.

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. We can obtain one depth map per two parallel images captured from the adjacent cameras, as shown in Fig. 2. Since there are 8 cameras in the KDDI-A test data set, seven depth maps are acquired for the first frame using a stereo matching process [7].

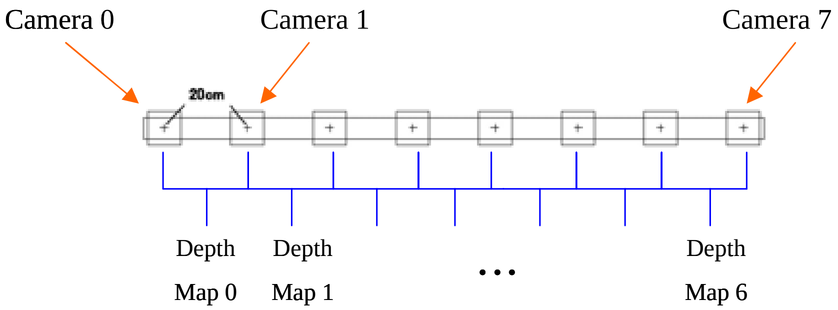
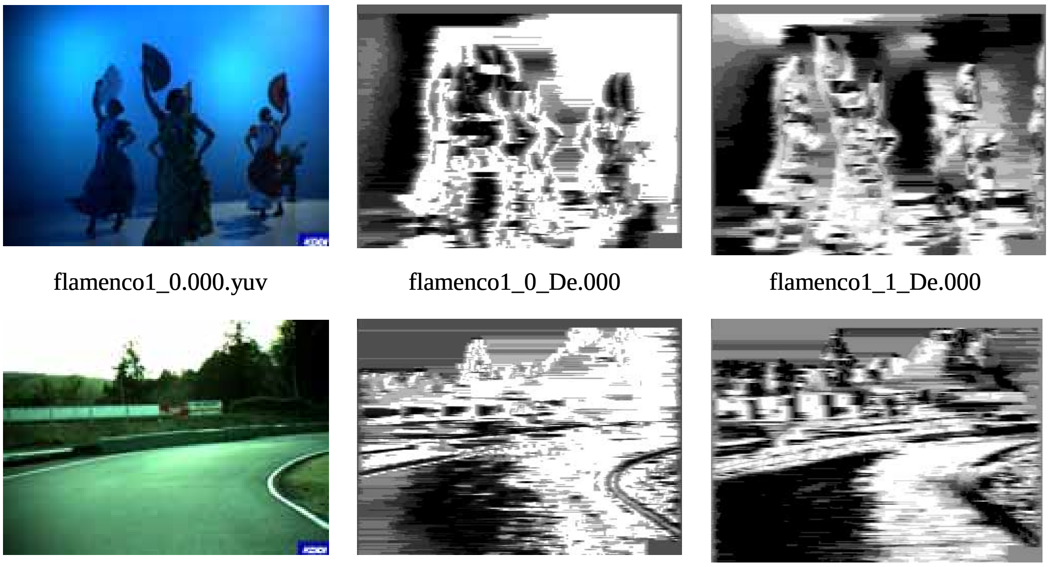


Fig. 2. Depth map extraction for KDDI-A test sequences.

There are several problems in generating LDI frames from the KDDI test sets. Although we can easily compute depth maps from disparity maps using the parallel configuration of cameras, quality of the computed depth map is not good. Figure 3 shows the histogram equalized depth maps obtained from the KDDI-A data sets: flamenco1 and race2. As shown in Fig. 3, it is hard to figure out depth positions of objects within the depth map. One of the main reasons is related to the image rectification that has already discussed in “Multi-view Video Coding using Image Stitching” [8].



flamenco1_0.000.yuv

flamenco1_0_De.000

flamenco1_1_De.000

race2_0.000.yuv

race2_0_De.000

race2_1_De.000

Fig. 3. Histogram equalized depth maps for KDDI-A data sets: flamenco1 and race2.

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 Eq. 1. It is not a simple task to obtain the disparity from the 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. There are two possible ways to get the disparity from the depth: one is rectification; the other is to compute the actual 3-D point location (x, y, z) in the world coordinate from the given depth. However, image rectification is an on-going research topic and we have not known perfect solutions yet.

Once we get LDI frames from the multi-view video, the inverse conversion is a default functionality of LDI; thus, it can be done easily without any complex procedure.

3 Framework for Multi-view Video Coding using LDI

Since MPEG-4 3DAV test sequences have been released, there are 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 sufficient additional information could be available soon. At the current stage, we propose a framework for encoding multi-view video using the concept of LDI.

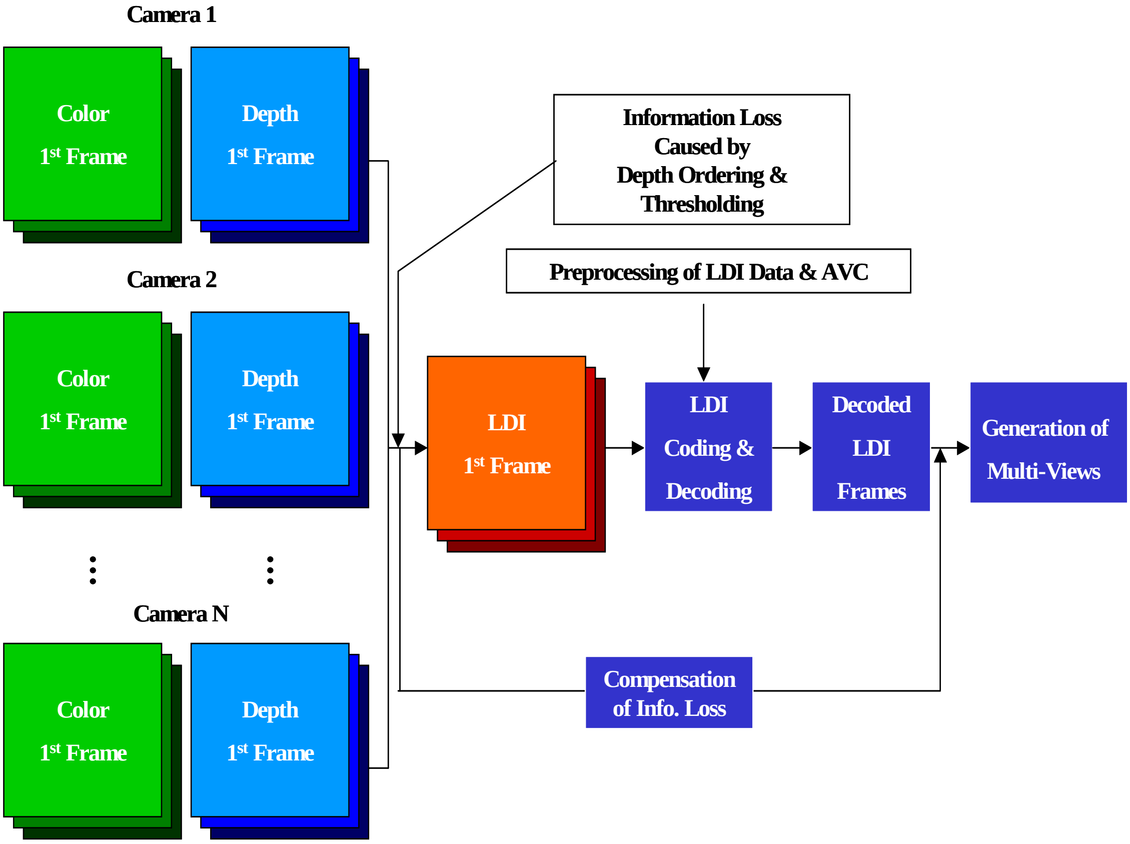


Fig. 4. A framework for multi-view video coding using the concept of LDI.

We obtain LDI frames from test multi-view video sequences by 3-D warping using given or acquired depth images. Since there is some information loss caused by depth comparison and thresholding in generating LDI, compensation is required before the reconstruction of multi-views. In the encoding step, LDI data are preprocessed [9] and AVC is applied to those processed data.

4 Conclusion

In this document, we have introduced a framework for encoding multi-view video using the concept of LDI. The proposed framework contains the conversion between multi-view video and LDI. The main idea is that multi-view video sequences with the depth information can be efficiently coded using the concept of LDI. We can apply efficient LDI coding schemes by constructing LDI frames from multi-view video with depth.

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

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

[7] A study on real-time extraction of depth and disparity map for multi-viewpoint images, ETRI Research Report, Nov. 2002.

[8] ISO/IEC JTC 1/SC 29/WG 11/M11292, "Multi-view Video Coding using Image Stitching," Oct. 2004.

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