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

ISO/IEC JTC1/SC29/WG11
MPEG2006/M13582
July 2006, Klagenfurt, Austria

Title: Complete Coding Result of Layered Depth Image Frames

Source: GIST
Authors: Yo-Sung Ho, Seung-Uk Yoon, Eun-Kyung Lee, 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). We have been proposed a framework for multi-view video coding using this concept of LDI as a 3-D approach unlike other 2-D based video coding techniques [1]. In this document, we describe how to encode the number of layers (NOL) and the residual information of LDI. In addition, we show the final coding results for LDI frames.

2 Coding of Number of Layers (NOL)

In our previous works [2][3][4][5], we have generated LDIs from the natural multi-view video sequence, e.g., "breakdancers". After generating LDI frames from the natural multi-view video with depth, we have separated each LDI frame into three components: color, depth, and the number of layers (NOL) as shown in Fig. 1 [5].

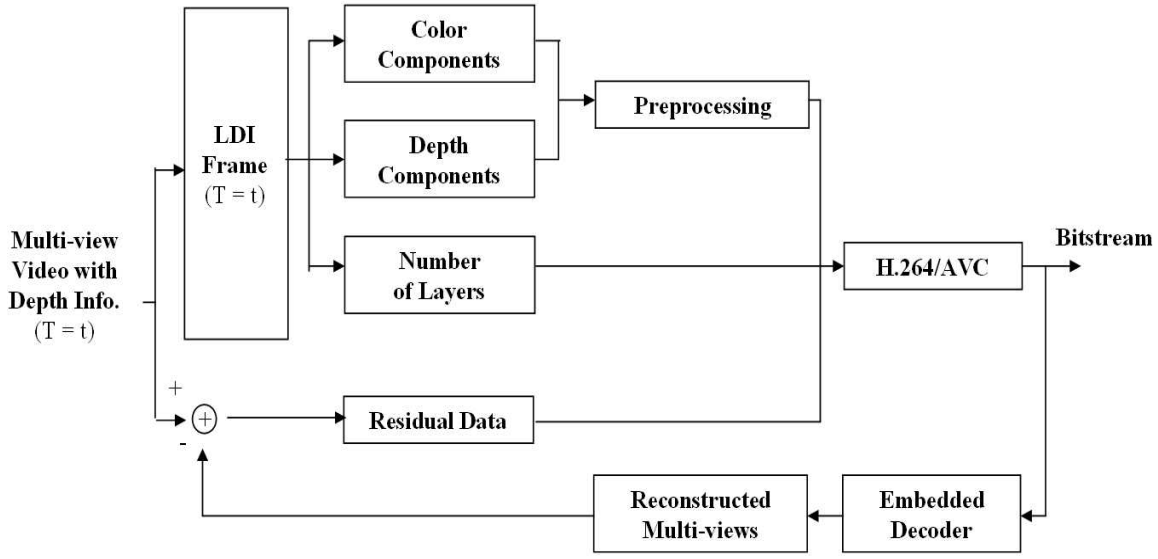

Fig. 1. Encoder structure for the coding of LDI frames.

For color and depth components, we have applied two kinds of preprocessing algorithms and observed that the layer filling technique have shown better performance in terms of saving coding bits [5]. Still remaining important data to be encoded are the NOL and the residual information. Therefore, we describe a method for the NOL coding and an algorithm to reduce the residual information in this document.

NOL could be considered as an image containing the number of layers at each pixel location. Figure 2 illustrates an example of the NOL image. Usually, the maximum number of layers is the same as the number of cameras used to capture the multi-view video. If we use eight cameras to acquire eight-view video, then the maximum number of layers is eight. The minimum number of layers is one because there always exists more than one layer. In other words, there are no empty pixels in the first layer of LDI.

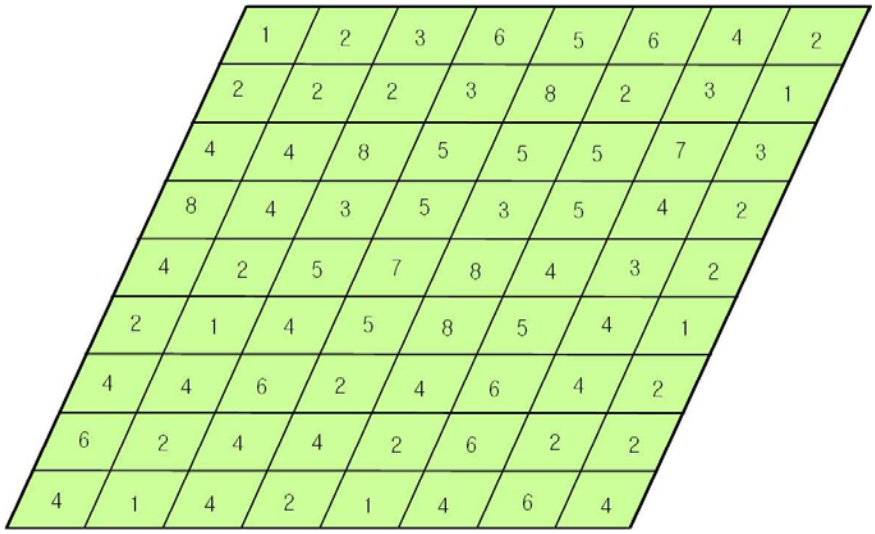

Fig. 2. The NOL image.

The physical meaning of NOL is that it represents the hierarchical structure of the constructed LDI in the spatial domain. If NOL is known, we can efficiently use empty pixel locations to manipulate the coherency between pixels. We can freely change the pixel orders, add dummy pixels in the empty locations, and remove them after the decoding process because we know where those pixels are.

Since the NOL information is very important to restore or reconstruct multi-view images from the decoded LDI, it is encoded by using the H.264/AVC intra mode. Since the dynamic range of the values of NOL is small, quantization noises can easily contaminate the reconstructed values. Consequently, it is difficult to restore the original NOL image.

In order to solve this problem, we change the dynamic range of the pixel values of the NOL image by considering both the encoding bits required for the changed dynamic range and the accuracy of restored NOL value.

$$\alpha \cdot nMinLayer \le \alpha \cdot V_{NOL} \le \alpha \cdot nMaxLayer, \;\; \alpha \le 255/V_{NOL} \;\; (\alpha \in N) \qquad (1)$$

where *nMinLayer* is the minimum number of layers, *nMaxLayer* is the maximum number of layers, V_{NOL} is the pixel value of the NOL image, and α is the scaling factor.

3 The Importance of the Residual Information

In the proposed encoder structure in Fig. 1, one of the most important parts is the residual data coding because it affects the overall compression efficiency. Currently, the residual information is used to fill out disoccluded area in the final reconstruction stage. Figure 3 shows the residual data extracted from the original multiple view images.

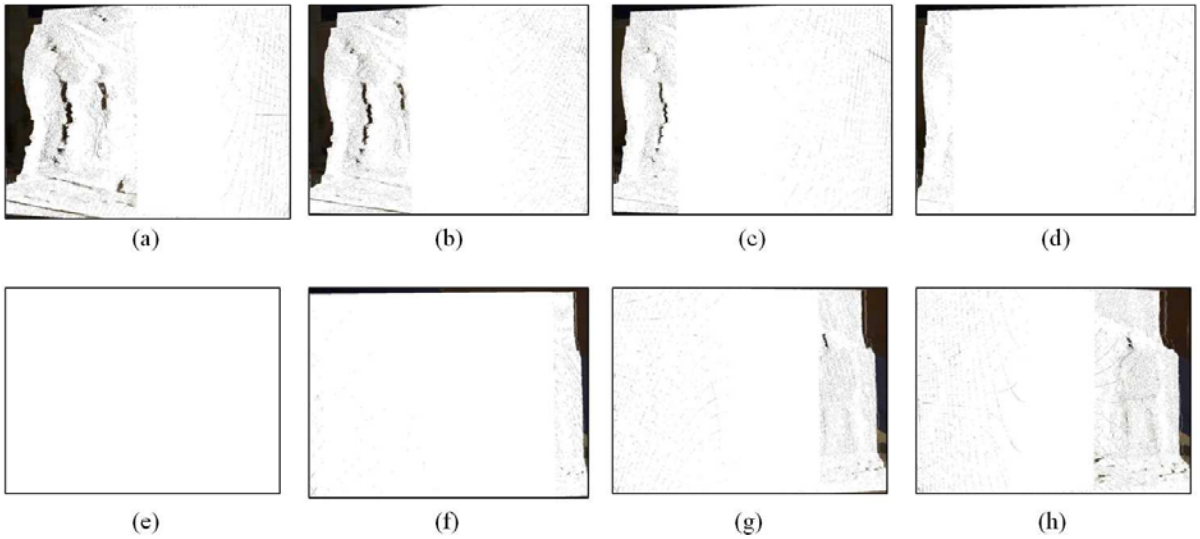

Fig. 3. Residual data: (a) view 0; (b) view 1; (c) view 2; (d) view 3; (e) view 4 (a reference view); (f) view 5; (g) view 6; (h) view 7.

Since the reference view is reconstructed almost perfectly, there is little residual information to be sent to the decoder as shown in Fig. 3(e). Most residual data are around left-most edges and right-most edges of those multi-view images. It seems that the residual data are relatively small amount compared to the main color or depth component, but more bits are consumed to encode them in practical experimentations.

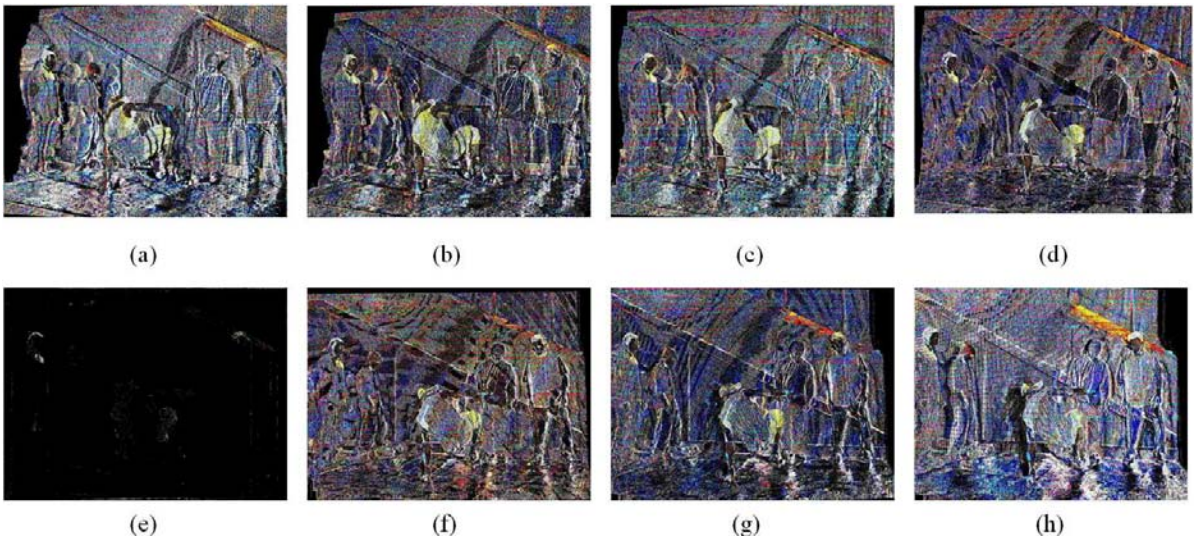

Fig. 4. Residual images: (a) view 0; (b) view 1; (c) view 2; (d) view 3; (e) view 4 (a reference view); (f) view 5; (g) view 6; (h) view 7.

Actual coding is performed for the residual image, not for the residual data. In other words, we calculate residual images after reconstructing final multiple view images using the residual data. Let us define a residual image is the differential image between the final reconstruction results and the original views. The residual data is additional information used to recover final multi-views as depicted in Fig. 3. Figure 4 shows residual images calculated by differentiating two images based on our definition. As we can observe from Fig. 4, these kinds of data consume lots of encoding bits, even more than the main components, in the current video codec, H.264/AVC. Therefore, it is essential to reduce the residual information to save bits for encoding.

4 Reduction of Residual Information using Pixel Interpolation

Theoretically, one way of reducing residual data is to reconstruct multi-views without using the information from the original images. It means that we should maximize the efficiency of using depth pixels (DPs) in back layers of LDI, neighboring pixels within a layer image, and spatial relationships between multiple images for the same scene.

In our reconstruction algorithms, there are three steps: the inverse 3-D warping; reconstruction without residual information; reconstruction with residual information [4]. In order to reduce residual information, we exploit the neighboring reconstructed pixels and images in our second reconstruction stage. As shown in Fig. 5, we can get intermediate reconstruction results after applying the inverse 3-D warping and depth ordering of the back layer pixels.

(a) (b) (c) (d)

Fig. 5. Reconstruction using back layers: (a) view 0; (b) view 1; (c) view 4; (d) view 7.

Our approach is to use the neighboring pixels and reconstructed images for interpolating empty pixels of the current reconstructed image. There are mainly two factors influencing the interpolation result: one is spatially located neighboring pixels within the current reconstructed image and the other is temporally located pixels in neighboring reconstructed images. We define the following equation to perform the pixel interpolation.

$$I_S(x,y) = \frac{1}{k} \cdot \sum_{i=0}^{W} \sum_{j=0}^{W} I(R_{(i,j)}) \tag{1}$$

$$I_T(x,y) = \sum_{n=0}^{N-1} a_n \cdot I(R_n), \quad \sum_{n=0}^{N-1} a_n = 1 \tag{2}$$

$$I_E(x,y) = \alpha \cdot I_S(x,y) + (1-\alpha) \cdot I_T(x,y), \quad 0 \le \alpha \le 1 \tag{3}$$

where $I_S(x,y)$, $I_T(x,y)$ is the intensity value of the interpolated pixel at the (x, y) position of the current image, $I_E(x,y)$ is the final interpolated pixel value, k is the valid number of

pixels within a W x W window, a_n and α are the weighting factors, and R means the reconstructed image. This equation is only applied for interpolating the empty pixels of the current image. The weighting factors have been determined by experiments.

Figure 6 shows the reconstruction results after performing the interpolation using the equation (3). We can observe that most holes except left-most and right-most sides are recovered with much less visual artifacts compared to the results in Fig. 5.

(a) (b) (c) (d)

Fig. 6. Reconstruction results using the pixel interpolation: (a) view 0; (b) view 1; (c) view 4; (d) view 7.

5 Encoding Result of LDI Frames

We have shown the comparison results in terms of PSNR Y vs. bitrate for the "breakdancers" sequence in Fig. 7. The proposed method is compared to others, which are extracted from the 75th MPEG documents. Since the rate control mechanism is not implemented yet for the LDI frames, we have manually adjusted total bitrates allocated for each LDI frame. For each LDI frame, four components, e.g., color, depth, NOL, and residual, are encoded by the proposed methods. We have calculated average bitrates per LDI frame and divided it by the number of views because each LDI frame hierarchically contains information for all viewpoints.

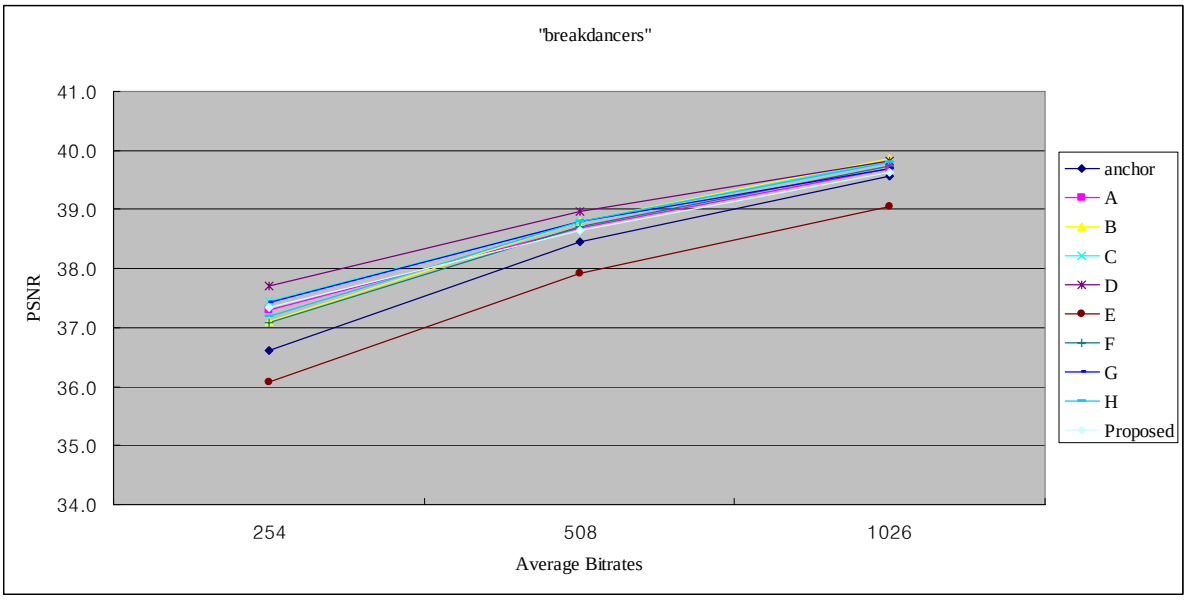

Fig. 7. PSNR Y vs. bitrate curves.

The proposed method has shown better performance than the anchor in terms of PSNR. Still remaining issues of the LDI-based approach are how to select the proper back layer pixels to fill out the current pixel location, how to dynamically adjust bitrates per each component, e.g., color, depth, NOL, and residual, and how to compare the performance of depth coding. Since the LDI frame explicitly contains the depth information and PSNR might not the best measure for evaluating the depth coding performance, it is needed to develop proper comparison metrics considering view generation results using the depth information.

6 Conclusion

In this document, we have explained the encoding method of the number of layers (NOL) and a pixel interpolation technique to reduce residual information of the layered depth image (LDI). NOL and residual data are very important components in encoding of the LDI because they affect the overall coding efficiency. Finally, we have shown the comparison results in terms of PSNR vs. bitrate for the temporal coding of LDI frames. From our past and current experiments, we believe that the proposed LDI-based framework could be a useful candidate for dealing with N-video plus N-depth data.

7 References

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

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

[3] ISO/IEC JTC1/SC29/WG11 m12485, “Generation and Coding of Layered Depth Images for Multi-view Video,” October 2005.

[4] ISO/IEC JTC1/SC29/WG11 m12849, “Reconstruction of Multi-view Images from Layered Depth Images,” October 2005.

[5] ISO/IEC JTC1/SC29/WG11 m13165, “Prediction Structures for the Constructed Layered Depth Image Frames,” April 2006.