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/M13165
April 2006, Montreux, Switzerland

Title: Prediction Structures for the Constructed Layered Depth Image Frames
Source: GIST and ETRI

Authors: Yo-Sung Ho, Seung-Uk Yoon, Sung-Yeol Kim, and Eun-Kyung Lee
(Gwangju Institute of Science and Technology)
Kugjin Yun, Sukhee Cho, and Namho Hur
(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). We have been proposed a framework for multi-view video coding using the concept of LDI as a 3-D approach unlike other 2-D based video coding techniques [1]. In this document, we describe the spatial and temporal prediction structure of the constructed LDI frames as the successive work of MVC using LDI [2][3][4].

2 Encoder Structure

In our previous works [2][3][4], we have generated LDIs from the natural multi-view video sequence, “Breakdancers”. The first eight color and depth frames of the 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 generating LDI frames from the natural multi-view video with depth, we separate each LDI frame into three components: color, depth, and the number of layers (NOL). Specifically, color and depth component consist of layer images, respectively. The maximum number of layer images is the same as the total number of views. In addition, residual data should be sent to the decoder in order to reconstruct multi-view images. Color and depth components are processed by data aggregation or layer filling to apply H.264/AVC. NOL could be considered as an image containing the number of layers at each pixel location. Usually, the range of the number of layers is dependant on the maximum number of cameras. 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. Finally, the residual data, differences between the input multi-view video and reconstructed ones, are encoded by H.264/AVC. Figure 2 shows the encoder block diagram. In Fig. 2, we have exploited two kinds of preprocessing methods to improve the spatial prediction efficiency of the LDI frame.

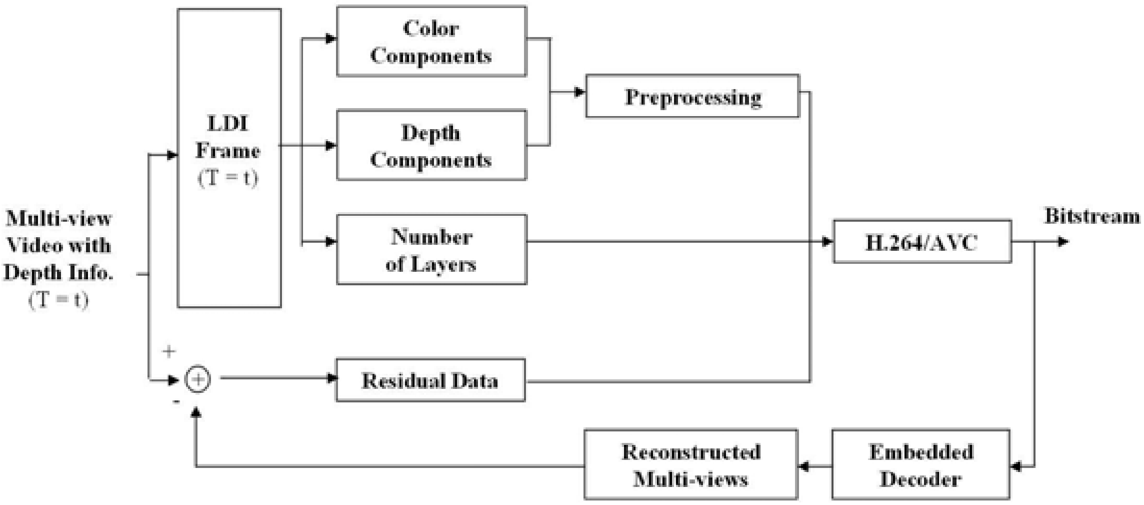


Fig. 2. Encoder block diagram

3 Spatial Prediction Structures within a Single LDI

In Fig. 2, the data aggregation [3] or layer filling is used in the preprocessing stage. The former was that we need to aggregate scattered pixels into the horizontal or vertical direction [4][5] because each layer of the constructed LDI has different number of pixels. Moreover, there are lots of empty pixel locations in back layers. Although H.264/AVC is powerful to encode rectangular images, it does not support shape-adaptive encoding modes. Therefore, we aggregate each layer image and then fill the empty locations with the last pixel value of the aggregated image. First, the scattered pixels in each layer are pushed to the horizontal direction. Second, the locally aggregated images are merged into a single huge image and empty pixels are padded. For example, if each layer image has XVGA (1024 x 768) resolution, the single aggregated image becomes 8192 x 768. Finally, the generated one is divided into the images with pre-defined resolutions, e.g., 1024 x 768, to employ H.264/AVC. One problem of the data aggregation is that the resultant images have severely different color distributions. It leads to poor coding efficiency because the prediction among aggregated images is difficult.

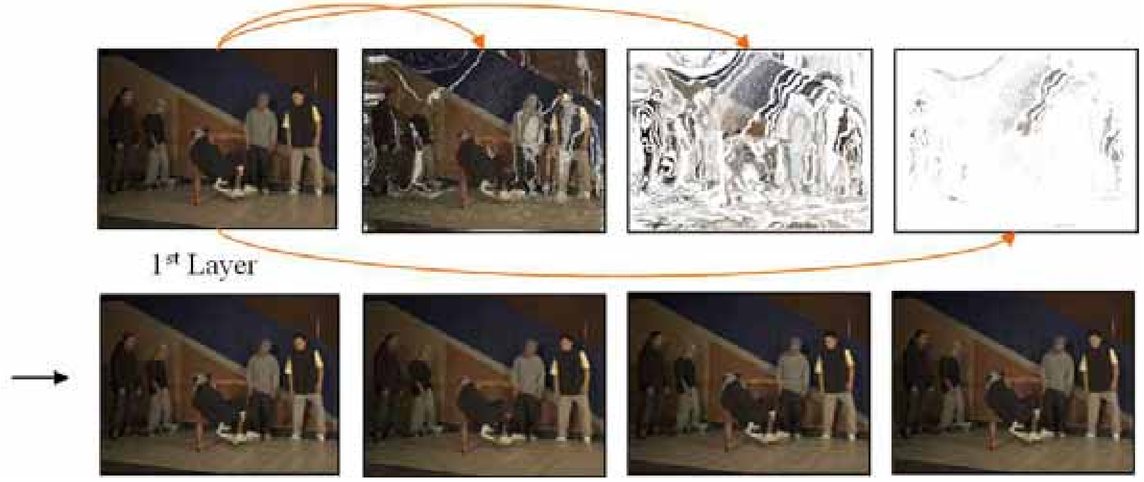


Fig. 3. The layer filling method for the spatial prediction within a single LDI frame

The second method is called the layer filling, which is shown in Fig. 3. In order to solve the above problem, we can fill the empty pixel locations of all layer images using pixels in the first layer. Since the first layer has no empty pixels, we can use same pixels in the first layer to fill the other layers. This increases the prediction accuracy of H.264/AVC, therefore data size could be reduced further. We can eliminate the newly filled pixels in the decoding process because the information of NOL is sent to the decoder. It is an eight bit gray scale image that each pixel contains an unsigned integer number representing how many layers there are. From our experiments, the layer filling is more efficient in the spatial prediction among layer images within a single LDI. The extended experimental result is listed in Table 1 as the depth threshold is changed. The NOL and residual data have not been contained in the Table 1.

Table 1. Data size comparison for the spatial prediction within a single LDI [Kbytes]

"Breakdancers"	1st 8 Frames	2nd 8 Frames
Sum of Frames (Color + Depth)	25,166	25,166
Simulcast (Color + Depth)	137.7	132.5
Simulcast (Color Only)	97.4	96.3
LDI Frame (Threshold: 0.0)	24,520	24,644
Encoded LDI (Aggregation)	135.4	133.7
Encode LDI (Layer Filling)	71.4	72.9
LDI Frame (Threshold: 3.0)	13,924	13,803
Encoded LDI (Aggregation)	131.7	133.8
Encode LDI (Layer Filling)	48.4	48.2
LDI Frame (Threshold: 5.0)	13,808	13,723
Encoded LDI (Aggregation)	91.7	93.0
Encode LDI (Layer Filling)	46.3	47.0

Figure 4 illustrates the PSNR result for each view using the constructed LDI frame with the depth threshold 3.0 at the fixed bitrate of 128 kbps. We expect that the PSNR result could be improved by adopting efficient interpolation methods because the result has been acquired without any interpolation or hole filling methods.

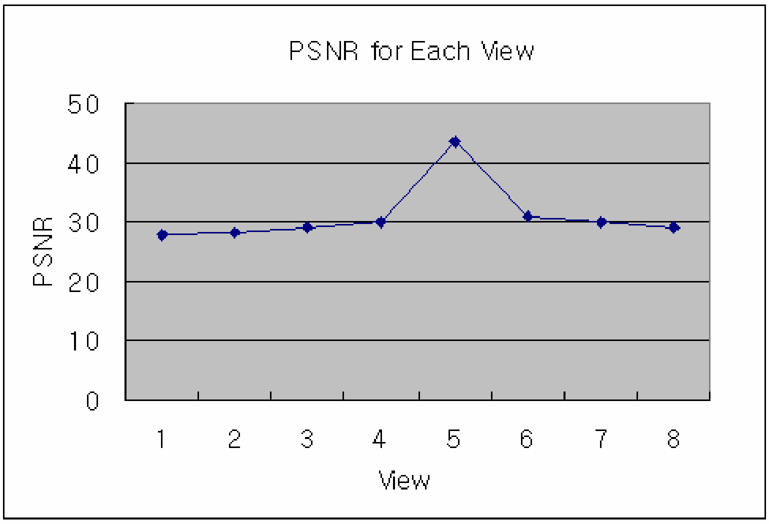
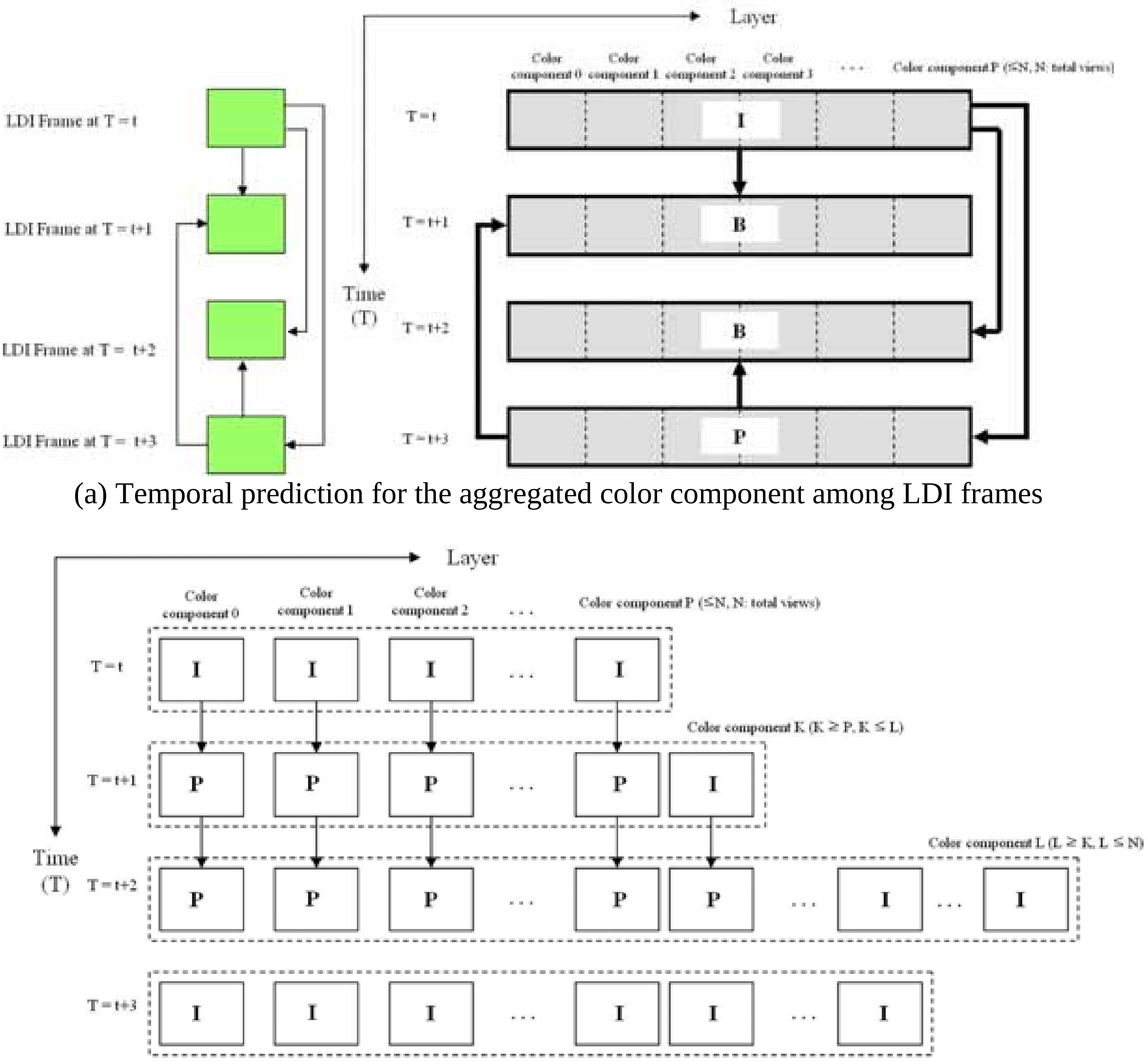


Fig. 4. PSNR chart for each view in the spatial prediction

4 Temporal Prediction Structures of the Constructed LDI Frames

In order to reduce the data size of multiple videos, the temporal prediction plays an important role rather than the spatial prediction among views [6]. Based on our framework for multi-view video coding [1], we need temporal prediction structures among the constructed LDI frames. In this document, we introduce two kinds of temporal prediction structures for color components of the constructed LDI frames as shown in Fig. 5.

Figure 5(a) shows the temporal prediction structure among LDI frames when we use the data aggregation in generating LDIs. It considers each LDI frame as a single image with a huge size. If we exploit the layer filling we do not need to perform data aggregation, then each color component could be predicted independently as depicted in Fig. 5(b). Figure 5(b) is an example because there could be existed different types of structures based on the number of layers for each LDI frame. For the depth component, we can apply IPIP… prediction structure because the depth information is very important for the reconstruction of multiple views.



(a) Temporal prediction for the aggregated color component among LDI frames

(b) Temporal prediction for the separated color component among LDI frames

Fig. 5. Example of temporal prediction structures among LDI frames

5 Conclusion

In this document, we have explained the encoder structure of the constructed layered depth image (LDI) frames. In addition, we have described spatial and temporal prediction structures based on our framework for multi-view video coding using the concept of LDI. For the spatial prediction of component images of the LDI, we have applied two kinds of approaches, e.g., the data aggregation and the layer filling. From our experiments, the layer filling shows better performance than the data aggregation. On the other hand, we have introduced temporal prediction structures among LDI frames. The first one is for the aggregated component of the LDI frame, and the other structure is for the separated component. We would perform more exploration and analysis for efficient prediction structures further based on our MVC framework.

6 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] J. Duan and J. Li, “Compression of LDI,” IEEE Transaction on Image Processing, Vol. 12, No. 3, pp. 365-372, March 2003.

[6] ISO/IEC JTC1/SC29/WG11 m12301, “Statistical Evaluation of Spatiotemporal Prediction for MVC,” July 2005.