# Three-dimensional Natural Video System based on Layered Representation of Depth Maps

Sung-Yeol Kim, Sang-Beom Lee and Yo-Sung Ho, *Member*, IEEE

**Abstract** — *In this paper, we propose a new three-dimensional (3-D) video system using a depth image-based representation (DIBR). In order to represent 3-D natural videos, we exploit texture images and synchronized depth maps, which are decomposed into four layers; number of layer (NOL), grid, boundary and feature point layers. With NOL, grid, and boundary layers, we define minimum required consecutive surfaces for 3-D natural videos. Then, we enhance their visual qualities by adding feature point layers. We regard NOL, grid and boundary layers as a base layer and feature point layers as enhancement layers to construct multi-layer representation of 3-D natural videos. With our proposed system, we cannot only represent depth maps efficiently, but also render 3-D natural videos in real time within reliable visual qualities. In addition, 3-D natural videos can be reconstructed adaptively according to consumers' capabilities and target applications within the framework of multi-layer representation. Experimental results have demonstrated that our proposed system can render and compress 3-D natural videos efficiently, as well as supporting the functionality of multi-view image generation.*[1]

**Index Terms — 3-D natural video system, depth image-based representation, layered representation of depth maps**

## I. INTRODUCTION

As rapid developments in the field of digital communications, computer powers, and high-speed networks, it is not too much to say that we are living in an age of the information revolution and the digital epoch. We are able to enjoy high quality audio-visual media with three-dimensional (3-D) displays and multi-channel speakers. Moreover, we can touch and manipulate provided contents directly with haptic devices [1]. In this wise, we are endeavoring to create reproductions, which are imitated with the real world, so as to feel the impression of 'being there', or *presence*, from them as much as possible [2].

Recently, high-definition (HD) videos have offered a wide field of view [3] and a 3-D TV has supported a natural viewing experience in true dimensions [4], [5]. Especially, 3-D natural videos play an important role to increase the feeling of *presence* as contents of 3-D TV. Basically, 3-D natural videos update 3-D surfaces constantly using texture images and their depth maps. One of main functionalities of 3-D natural videos is to generate multi-view images, which enable consumers to experience more realistic visual scenes [6].

In general, multi-view image generation can be accomplished by two kinds of systems; multi-view camera systems [7] and depth image-based rendering (DIBR) [8], [9] systems. Multi-view camera systems employ a number of cameras to obtain wide-viewing angle images. As displaying captured multi-view images according to consumers' viewpoints, we can be immersed in the content naturally. However, a multi-view camera system requires complicated coding and transmission techniques in proportion to the number of cameras so as to deal with their data within a limited bandwidth channel.

On the other hand, DIBR systems make use of two kinds of images; texture images and synchronized depth maps. With DIBR techniques, we are able to render arbitrary views from 3-D scenes reconstructed by two input sources. Although DIBR systems can support narrow-viewing angle images, they are considered as a suitable main theme for 3-D natural video system, since we only need to process two streams to support the functionality of multi-view image generation.

Basically, depth maps include depth values of 256 levels and are provided as gray-level images. Depth values in a depth map indicate the depth information of pixel positions in the corresponding texture image. Depth data can be acquired by a depth camera directly [10], or extracted by a stereo matching algorithm [11], [12]. In a depth map, we can regard pixel positions and depth values as geometry information in the 3-D domain. Moreover, we can think of a texture image as texture data, which cover a geometric surface generated by geometry information extracted from its depth map.

There are two approaches in 3-D natural video systems using DIBR techniques; mesh-based depth representation [13], [14] and image-based reconstruction [15]. In the mesh-based depth representation, we extract feature points in a depth map and generate a 3-D surface using a mesh triangulation technique [16], [17] as shown in Fig. 1. The main advantage of mesh-based depth representation is a high rendering speed, since we usually exploit a graphic accelerator to render 3-D consecutive scenes. However, we suffer from dealing with its data due to its irregularities.

In the image-based reconstruction, we regard depth maps as 2-D images. Therefore, we use all pixel data in a depth map to display 3-D surfaces. However, although the data of an image-based reconstruction are easy to deal with, we need more rendering time and reasonable hole-filling techniques.

[1] This research was supported by the MIC, Korea, under the ITRC support program supervised by IITA (IITA-2005-C1090-0502-0022) through the Realistic Broadcasting Research Center (RBRC) at the Gwangju Institute of Science and Technology (GIST).

S.-Y. Kim is with the Department of Information and Communications, GIST (e-mail: sykim75@gist.ac.kr)

S.-B. Lee is with the Department of Information and Communications, GIST (e-mail: sblee@gist.ac.kr)

Y.-S. Ho is with the Department of Information and Communications, GIST (e-mail: hoyo@gist.ac.kr)

Contributed Paper
Manuscript received June 29, 2006

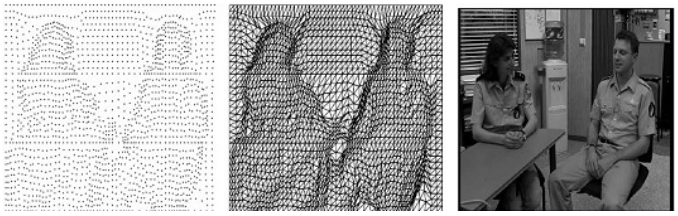


**Fig. 1. Mesh-based depth representation of DIBR for a 3-D video system**

In this paper, we focus on the mesh-based depth representation of DIBR for a 3-D natural video system with preservation of regularities like an image-based reconstruction. We introduce a layered representation of depth maps and render 3-D consecutive scenes progressively within the framework of multi-layer representation.

This paper is organized as follows. In Section II, we present the architecture of a new 3-D natural video system briefly. Then, Section III explains a layered representation of depth maps and Section IV describes the functionality of multi-layer representation of the proposed system. After providing experimental results in Section V, we conclude in Section VI.

## II. System Architecture

In this paper, we propose a new 3-D natural video system using a layered representation of depth maps. While current video systems deal with 2-D audio-visual media and only support simple interactions, the proposed system provides a hierarchical structure to express 3-D natural videos and supports user-friendly functionalities. Figure 2 shows the overall system architecture of the proposed video system.

The proposed system can be divided into two parts; a sender and a receiver. First, texture images and their depth maps are obtained from one of two systems at a sender side, either a depth camera system or a multi-view camera system. When we use a multi-view camera system, we obtain depth data using a stereo matching technique. Next, depth maps are smoothed to generate multi-view images without geometric distortions. After decomposing depth maps into four layers; number of layer (NOL), grid, object boundary, and feature point layers, we compress them by a video codec, H.264/AVC coder [18], with texture images. Before compressing texture images, we convert them into squared images to be used for texture mapping. Finally, bit streams of 3-D natural videos are transmitted to a receiver through a channel.

At a receiver side, transmitted bit streams will be decoded so as to recover depth maps and squared texture images. After generating a hierarchical structure with four layers of depth maps, we synchronize them with received texture images. In order to construct multi-layer representation for 3-D natural videos, we regard NOL, grid and boundary layers as a base layer, and feature point layers as enhancement layers. According to consumers' capabilities and target applications, a layer selector determines the amount of required data to represent 3-D natural videos. Finally, we can provide consumers with 3-D surfaces constantly with multi-view images or stereoscopic images [19].

Our proposed system is different from previous multi-layer representation methods such as progressive meshes (PM) [20] and layered depth images (LDI) [21]. First, although PM is similar to the framework of multi-layer representation for mesh-based video scenes, it does not use depth maps and cannot decompose them into layered images structurally. Also, our system is different from LDI, which generates multi-view images using layered representation of multiple depth maps. LDI uses NOL like our system. However, LDI is an image-based rendering technique basically. On the other hand, the proposed system is a mesh-based depth representation.

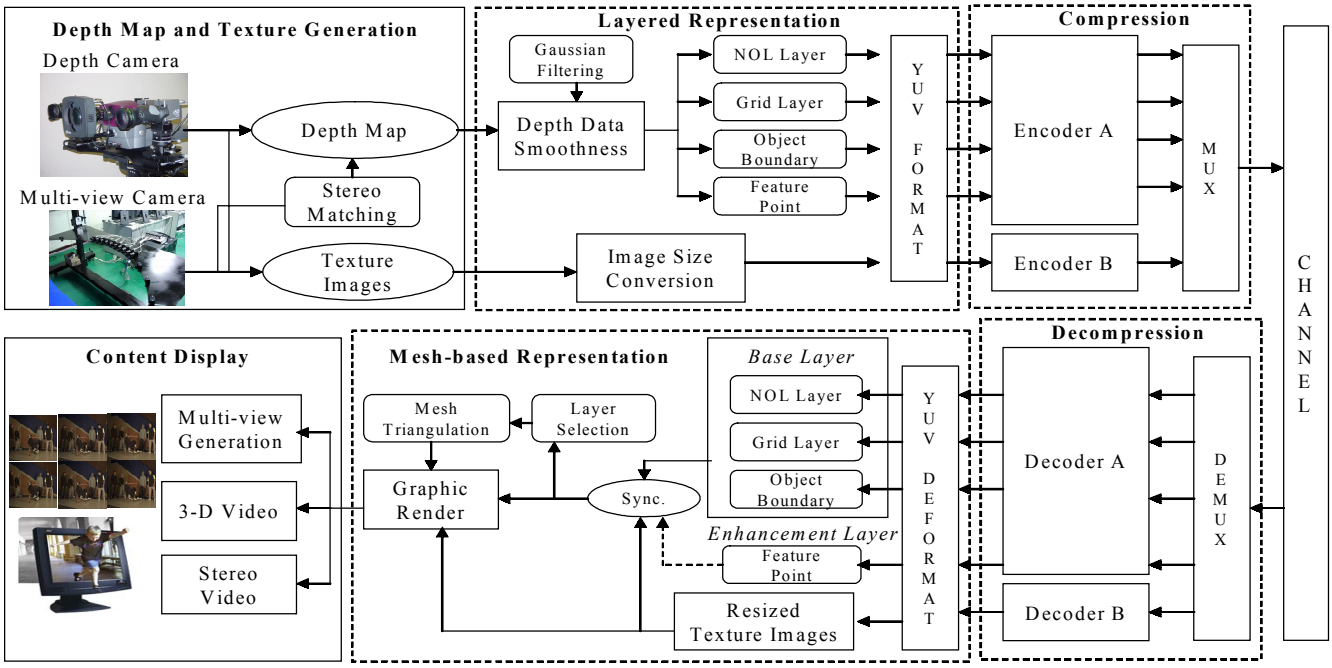



**Fig. 2. Overall system architecture of the proposed 3-D natural video system**

## III. LAYERED REPRESENTATION OF DEPTH MAPS

In order to represent 3-D natural videos, we express depth maps with a layered representation, which includes grid, boundary, NOL and feature point layers. With the layered representation, we can offer the convenience to deal with the data of mesh-based depth representation and simple structure to generate 3-D consecutive natural scenes.

### A. Grid Layer

We define a 3-D initial surface with a depth map for each frame, called as a grid layer. Grid layers can be generated by carrying out down-sampling with a depth map according to the size of grid cells, which is the unit of a grid layer. In this paper, we define the size of grid cell as $2^m \times 2^n$ pixel resolutions, such as 16×16, 8×8, or 16×8, since we need to minimize distortions in the initial surface obtained from a grid layer and maintain the property of regularities.

With a grid layer, we find out 3-D geometry data to construct an initial surface and then make mesh surfaces using a mesh triangulation technique. In order to make a grid layer, we divide a depth map into blocks by the size of a grid cell. First, we find four corner pixels with counter-clockwise direction in a grid cell. Since we can regard four corner pixels as four vertices, which have $x$, $y$ and $z$ coordinates in the 3-D space, we can define these vertices, $v_1$, $v_2$, $v_3$ and $v_4$, by following definitions.

$$v_1=\{x_1, y_1, z_1\},\ v_2=\{x_2, y_2, z_2\},\ v_3=\{x_3, y_3, z_3\},v_4=\{x_4, y_4, z_4\} \quad (1)$$

With four vertices, we construct two triangles. One is composed of $v_1$, $v_2$, and $v_3$. The other is composed of $v_1$, $v_3$, and $v_4$. Also, we can generate texture coordinates with four vertices if we know the depth map resolution. When there is a depth map with $l \times m$ image resolution, texture coordinates, $t_1$, $t_2$, $t_3$, and $t_4$, will be fallowing Eq. 2.

$$t_1=\{x_1/l, y_1/m\},\ t_2=\{x_2/l, y_2/m\},\ t_3=\{x_3/l, y_3/m\},\ t_4=\{x_4/l, y_4/m\} \quad (2)$$

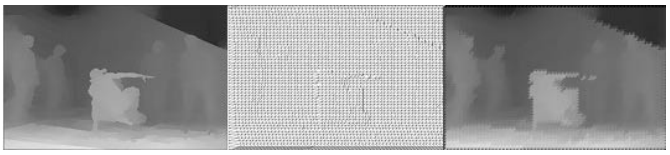
If the size of a grid cell is p × q pixels, the total number of grid cells will be like Eq. 3.

$$GC_{num} = l/p + m/q \quad (3)$$

Also, the total number of triangles will be like Eq. 4.

$$T_{num} = 2\times(l/p + m/q) \quad (4)$$

As a result, we obtain a down-sampled image with $l/p + m/q$ resolution as a grid layer.



(a) Original depth map (b) Wire-frame of a grid (c) Depth map of a grid

**Fig. 3. Initial surface generation with a grid layer**

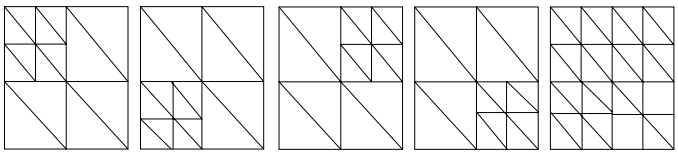
The main advantage that we employ a grid layer in layered representation of depth maps is to support regularities like 2-D videos. As a matter of fact, it has been hard to define what the minimum data can represent initial surfaces of 3-D natural videos, since we should find out regularities from input sources. Previous mesh-based depth representation methods in DIBR could not support such regularities so that it was difficult to deal with their data. However, we extract regularities easily with a depth map in this paper, since we can divide it uniformly using its property like general 2-D images. As shown in Fig. 3, we note that a grid layer represents an initial surface of 3-D natural videos successfully. However, it is not enough to provide consumers with high-quality visual services. Therefore, we need to enhance the initial surface with additional information.

### B. Boundary Layer

In order to improve visual qualities of initial surfaces generated by a grid layer, we insert a boundary layer. In general, we need to deal with object boundaries carefully since serious distortions are mainly occurred in their areas [11]. There are mismatches between object boundaries and texture segments. In other words, some parts in texture segments can be the region of the object boundary and the others can be the region of background.

In this paper, we employ a quad-tree structure and a full modeling technique adaptively to maintain object boundaries. In a quad-tree structure, we partition a grid cell consecutively until the background regions are removed. First, we determine whether a grid cell is included by the region of an object boundary or not using an edge detection algorithm, such as Sobel filtering. Then, we partition a grid cell into four sub-grid cells. When partitioned sub-grid cells are still in the region of an object boundary, we repartition them constantly. Finally, we can represent initial surfaces of 3-D natural videos more accurately and improve visual qualities.

According to the region of object boundaries, we also employ a full modeling technique. When the region of object boundary occupies more than a half of a grid cell, we use all pixel information in the grid cell. As a result, there are five types in a boundary layer as shown in Fig. 4.



(a) Type 1 (b) Type 2 (c) Type 3 (d) Type 4 (e) Type 5

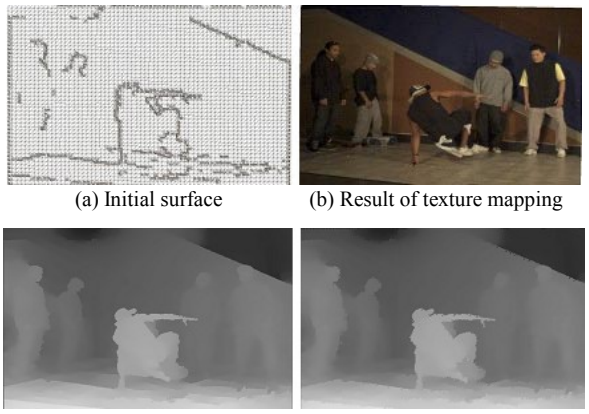
**Fig. 4. Types of a boundary layer**

A quad-tree structure is described from *type 1* to *type 4* for a boundary layer. In a quad-tree structure, we need 13 depth pixels to describe a grid cell. In case of *type 5* for a boundary layer, i.e. a full modeling technique, there will be 16 depth pixels in a grid cell. Therefore, when we know the type and its depth information in advance, we can construct the shape of a boundary layer.

With five types and depth information, a boundary layer can be converted into two images; one is for indicating the type of object boundary and the other is for notifying the corresponding depth information. These two images have the same resolution with a grid layer. As a matter of fact, the former image can be a part of NOL, which will be introduced in the following section. Therefore, we reconstruct minimum

required surfaces for 3-D natural videos with three images; grid and boundary layers.

Figure 5 shows the result of the surfaces with grid and object boundary layers. Although some distortions are occurred in the object boundary regions, we can notice that a depth map generated by initial surfaces is similar with the original depth map. As a result, we can reconstruct 3-D consecutive initial scenes successfully.



(a) Initial surface (b) Result of texture mapping

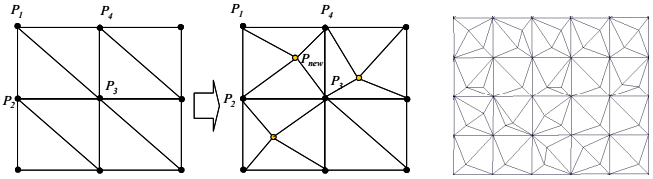
(c) Original depth map (d) Depth map of initial surfaces

**Fig. 5. Enhanced initial surfaces with boundary layers**

## C. Feature Point Layer

After generating minimum required surfaces for 3-D natural video with gird and boundary layers, we need to enhance 3-D surfaces with additional information since it is not enough to provide consumers with more high-quality visual services. In general, the region generated by a grid layer has low visual quality. As a result, we need to insert more information in the surface generated by a grid layer.

In this paper, we employ feature point layers to improve visual qualities. Feature points mean depth pixels in a grid cell affecting the shape of surfaces. Therefore, we first need to select one which has a great influence on the shape of surfaces. Then, we select the next one according the influence on the shape of surfaces. In principal, we can select feature points as the same as the total number of pixels in a grid cell except grid layers. However, two or three feature points usually exist in a grid cell since regions in a grid layer are continuous geometrically. Figure 6 shows feature points in grid cells.




(a) Grid layer (b) Feature point (c) Reconstruction

**Fig. 6. Feature point layers**

After receiving feature points in a grid cell, we regenerate more complex surfaces with Delaunay triangulation [19] at a receiver side. When we extract the feature points at a sender, we exploit a modified maximum distance algorithm so as to keep the optimal manner. Figure 7 shows the algorithm to extract feature points.

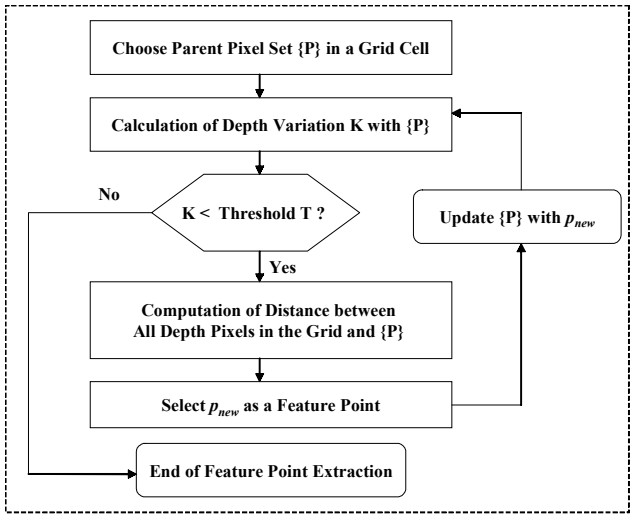



**Fig. 7. Feature point extraction**

In order to calculate the distance between all depth pixels in a grid cell and parent pixel {P}, we use Eq. 5,

$$D_i = \frac{1}{n\{P\}} \sum_{j=0}^{n\{P\}} (x_i - p_j)^2, (0 \le i < N) \qquad (5)$$

where, $x_i$ and $p_j$ mean the current depth pixel and the parent pixel, respectively. $n\{P\}$ is the total number of parent pixels.

When the number of feature points is *k,* feature point layers can be converted into *k+1* images with corresponding depth information. These images have the same resolution of one for a grid layer. One is for indicating the number of feature points and the others are for notifying the corresponding depth information. As a matter of fact, the former image can be a part of NOL, which will be introduced in the next section. Therefore, we reconstruct enhanced surfaces for 3-D natural videos with feature point images. Figure 8 shows the reconstruct result of 3-D videos with feature point layers. We notice that a depth map generated by 3-D surfaces with feature point layers is almost the same as the original depth map.

(a) Wire-frame (b) Texture mapping (c) Depth map

**Fig. 8. Reconstruction of 3-D natural videos with feature point layers**

## D. Number of Layer

In this paper, we employ NOL layers to construct the 3-D video structure with grid, boundary, and feature point layers. NOL is important data in the proposed system, since we recognize the number of feature points, the type of a boundary layer, and the position of a grid layer with NOL. Naturally, there will be a NOL for each frame. Figure 9 explains well what NOL layers play a role in. The left side picture indicates a NOL layer and the right side picture shows the structure of a 3-D scene with the NOL layer.

When the maximum number of feature points is *n*, the type of boundary layer will be from *n*+1 to *n*+5. A quad-tree structure of a boundary layer will be matched with one of

from $n$+1 to $n$+4. Also, The NOL of a full modeling in a boundary layer will be $n$+5. As a result, when the NOL in a grid cell is $n$, there will be a grid layer and $n$ feature point layers. On the other hand, when the NOL is $n$+5, there will be a grid layer and a boundary layer with a full modeling option. Naturally, when NOL in a grid cell is zero, there will be only a grid layer.

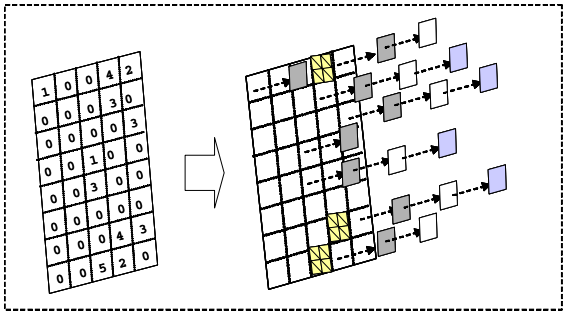


**Fig. 9. Representation of a layered structure with NOL layer**

The maximum number of feature points in the example is three. Therefore, the NOL of a boundary layer can be one of from four to nine and the maximum number of NOL is nine.

## IV. MULTI-LAYER REPRESENTATION

### A. Structure of a 3-D Natural Frame

Our proposed system is a kind of multi-layer structure, which has a base layer and several enhancement layers for each frame. In this paper, we regard a base layer as the combination of NOL, grid, and boundary layers. Also, we regard enhancement layers as feature point layers. Such a multi-layer representation enables consumers to render 3-D natural videos adaptively according to users' capabilities or multimedia platforms.

With a base layer, we can support minimum data to render 3-D natural videos. When we want to display 3-D natural videos within a small multimedia platform, such as a cellular phone or a PDA, we need not to use whole structure of our proposed system. On the other hand, when we want to show 3-D natural videos within a personal computer or a television system, we have to employ enhancement layers to support high-quality visual services. In order to use grid and boundary layers in a base layer, we just ignore the NOL number for feature point layers. Otherwise, we activate all NOL numbers and exploit feature point layers.

When a depth map has $l \times m$ resolution and the size of grid cells is $p \times p$ pixels, we decompose it into four layer images that have $(l/p+1) \times (m/p+1)$ resolutions as mentioned in Section III. In addition, we covert texture image sequences into square ones having $2^n \times 2^n$ resolutions to carry out texture mapping. As a result, we render 3-D natural videos with four layer streams and square texture image sequences.

In this point, we should note that our proposed system provides regularities like 2-D images in sprit of mesh-based depth representation. That is, we do not send 3-D data itself of mesh-based depth representation to a receiver but send four layer images and texture sequences. Figure 10 shows the concept of multi-layer representation of our system.

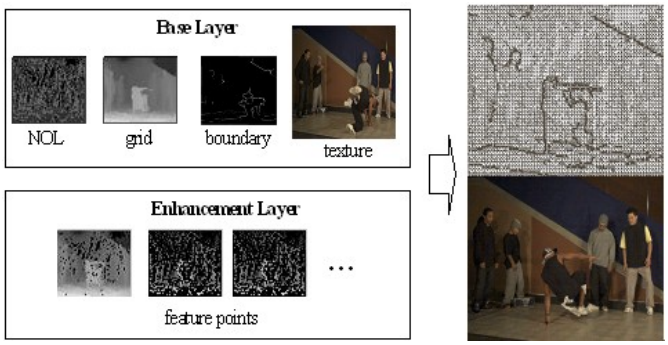



**Fig. 10. Multi-layer representation**

In our system, we can render 3-D dynamic scenes progressively in consideration with transmitted environments. In addition, the proposed structure provides us with systemic convenience for some 3-D processing techniques, such as level-of-detail (LOD), compression, and visual scalability.

### B. Compression of Depth Maps

In order to code depth maps and texture images, we employ a H.264/AVC coder. First, NOL sequences are coded within the lossless manner since they are the most important data in our structure to represent 3-D natural videos. When we compress NOL sequences, we multiply an integer $k$ to each NOL before converting them into a YUV stream. The multiplier $k$ will be determined by the quantization step size in a H.264/AVC coder. In order to code grid, object boundary, and feature point image sequences, we also use a H.264/AVC coder. There are no holes in a grid image. However, there are lots of holes in the object boundary and feature point images so that we need to fill out them toward increasing compression ratio. We can use average values for filling holes.

### C. Measurement

In order to evaluate the visual quality of our system, we use a modified Hausdroff distance. Basically, Hausdroff distance [22] is widely used to measure the distortion among 3-D mesh models. The modified Hausdroff distance evaluates the distortions between a reconstructed 3-D scene and a surface including all depth pixels in the corresponding depth map. Especially, we only need to consider the depth differences between them. The closest depth pixel in a reconstructed 3-D scene with a depth pixel in a depth map can be selected with the definition of distance by Eq. 6,

$$d(p,S') = \min_{p' \in S'} \| p - p' \| \tag{6}$$

where $d(p,S')$ denotes a distance between the reconstructed surface $S'$ and depth pixel $p$ in the depth image. The term of $P'$ indicates the depth pixel in $S'$. As a result, the modified Hausdroff distance can be defined by Eq. 7.

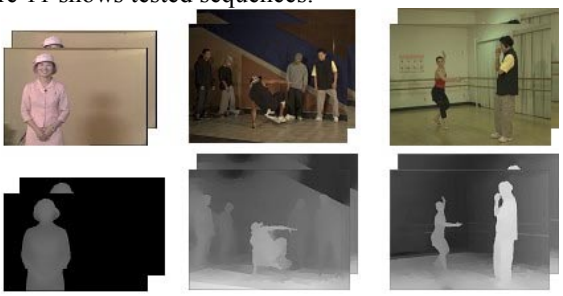
$$d(S,S') = \max_{p \in S} d_z(p,S') \tag{7}$$

Here, S denotes the surface including all depth pixels in a depth image and $d_z(p,S')$ indicates the $z$ coordinate distance between the selected depth pixel and the depth pixel $p'$ in $S'$.

## V. EXPERIMENTAL RESULT

We tested the performance of our system with several depth maps and texture image sequences. Tested sequences were Home-shopping, Break-dance and Ballet. Home-shopping sequences have 100 frames with 720×486 resolutions, while Break-dance and Ballet sequences have 100 frames with 1024×768 resolutions. Depth maps for Home-shopping are captured by a depth camera, ZCam$^{TM}$ [23]. Other depth maps are obtained by a stereo matching algorithm.

Home-shopping was made so as to use as broadcasting contents for 3-D TV at Realistic Broadcasting Research Center (RBRC) in Korea [24]. Break-dance and Ballet sequences were provided by Microsoft as test sequences for multi-view video coding (MVC) in MPEG standard [11]. Figure 11 shows tested sequences.

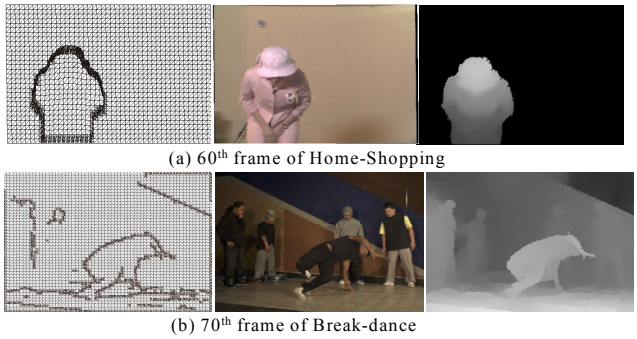


(a) Home-shopping (b) Break-dance (c) Ballet

**Fig. 11. Test sequences**

### A. Generation of Multi-layer Representation

In order to generate grid layers, we used 16×16 pixels as the size of grid cell. For all test sequences, we generated a base layer and two enhancement layers. Figure 12 shows the result of base layers of 60$^{th}$, 70$^{th}$, and 61$^{st}$ frame of test sequences in an order, respectively. As we mentioned in Section IV, we constructed minimum required surfaces of 3-D natural videos with NOL, grid, and boundary layers. As shown in Fig. 12, we could maintain the reliable visual qualities with a base layer when we compared with a full modeling using all depth data in a depth map. The main reason that we could maintain visual qualities was that we processed the region of boundaries carefully using a quad-tree structure and a full modeling technique adaptively.



(a) 60$^{th}$ frame of Home-Shopping

(b) 70$^{th}$ frame of Break-dance

(c) 61$^{st}$ frame of Ballet

**Fig. 12. Results of base layers**

Figure 13 and Figure 14 show the result of enhancement layers for 60$^{th}$, 70$^{th}$ and 61$^{st}$ frames of test sequences. We added enhancement layers to improve visual qualities from a base layer. In Fig. 13, we inserted one enhancement layer, i.e. a feature point per a grid cell. We could notice that the visual qualities of 3-D natural videos improved better than ones with a base layer in Fig. 12. Furthermore, we could notice that the surfaces reconstruct by a base layer and two enhancement layers in Fig. 14 were more complex and better visual quality than ones reconstructed by a base layer and an enhancement layer in Fig. 13.

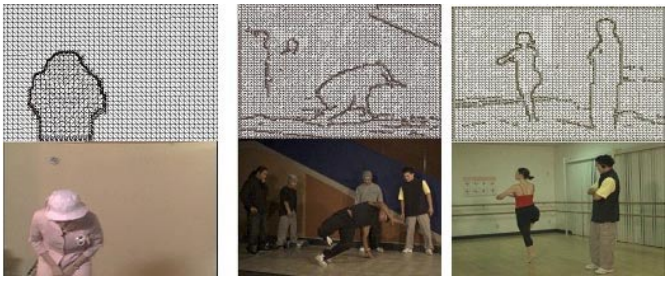


**Fig. 13. Results of base and one enhancement layers**

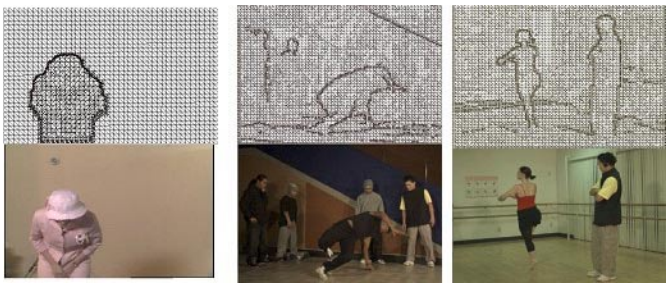
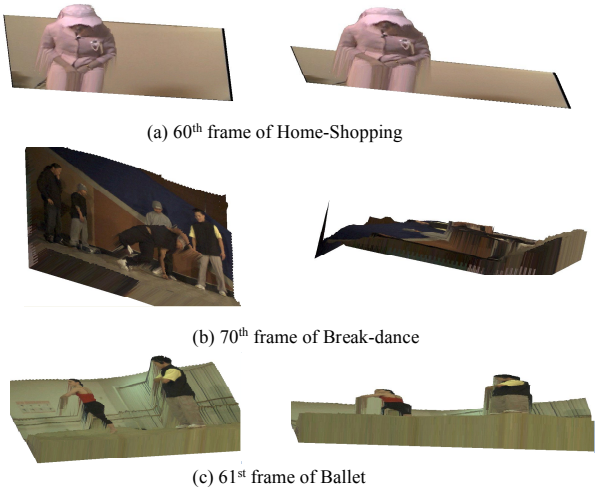


**Fig. 14. Results of base and two enhancement layers**

As a result, we could render 3-D natural videos progressively with reliable visual qualities with such a multi-layer representation of depth maps. In order to support the content using 3-D natural videos to consumers within the limited channel capacities, we could control the number of enhancement layers in proportion to required visual qualities related to target multimedia platforms. Figure 15 shows the result of reconstructed surfaces with a base layer and two enhancement layers.



(a) 60$^{th}$ frame of Home-Shopping

(b) 70$^{th}$ frame of Break-dance

(c) 61$^{st}$ frame of Ballet

**Fig. 15. Results of reconstructed surfaces**

We also compared the visual quality of reconstructed scenes with full modeling and our system. As shown in Fig. 16, we used modified Hausdroff distance to measure the distortion. We could notice that the more enhancement layers we transmitted, the higher visual quality we could support.

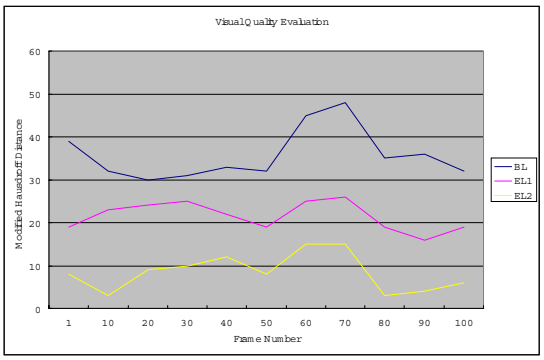



**Fig. 16. Evaluation of visual qualities**

In addition, we could render 3-D consecutive surfaces in real time with mesh-based reconstruction. Table 1 shows the comparison result of rendering time with a full modeling using all depth data and our scheme for Home-shopping. With our system, we could render 3-D natural video in real time.

**TABLE 1**
**RENDERING TIME ANALYSIS**

| Test Sequence | | Num. of Tri. | Render time (sec.) | Frame rates (f./sec.) |
|---|---|---|---|---|
| Home-shopping | Full modeling | 344,922 | 1.532 | 0.65 |
| | Only Base | 3202 | 0.027. | 37.04 |
| | Base + EL 1 | 13802 | 0.065 | 18.18 |
| | Base + EL 2 | 14204 | 0.068 | 12.82. |

## B. *Compression of Depth Maps*

We compressed depth maps for Home-shopping and Break-dance sequences using H.264/AVC video coder unlike previous mesh-based depth representation as we mentioned in Section IV. Home-shopping was composed of depth maps with small depth variation. On the other hand, Break-dance consisted of depth maps with larger depth variation than Home-shopping. Previous works usually used a 3-D mesh coder to compress the data of 3-D scenes. Consequently, they suffered from irregularities of 3-D data. However, we could exploit a video codec directly since our scheme supported regularities structurally. Table 2 shows the compression results of our scheme and original depth maps.

**TABLE 2**
**COMPARISON OF COMPRESSION RESULT FOR DEPTH MAPS**

| Test Seq. | QP | Original Maps | | Proposed Scheme | |
|---|---|---|---|---|---|
| | | Bits | PSNR | Bits | PSNR |
| Home-Shopping | 30 | 726,104 | 45.02 | 287,168 | 44.07 |
| Break-Dance | 30 | 3,764,616 | 42.47 | 1,617,152 | 42.22 |

## C. *Multi-view Image Generation*

In order to generate multi-view images, we rotated a virtual camera around reconstructed 3-D surfaces from -15 to +15 degrees from the center viewpoint. Figure 17 shows results of multi-view image generation with a virtual camera rotated -15, -10, and -5 degree without preprocessing of depth maps. As shown in Fig. 16, when we generate multi-view images without preprocessing of depth maps, we met geometric distortions, called as *rubber-sheet artifacts* due to disocclusion areas appearing at virtual viewpoints [25].

(a) Result of multi-view generation (b) Rubber-sheet artifacts

**Fig. 17. Multi-view generation without preprocessing of depth maps**

In order to prevent rubber-sheet artifacts, we employed an adaptive Gaussian smoothing filter. After extracting object boundaries from the depth map using a Sobel filter, we applied smoothing filters with various window sizes according to the amount of depth variations. As a result, we removed the geometric distortions efficiently when we made multi-view images with depth maps and texture images. Figure 18 shows results of multi-view image generation with a virtual camera rotated -15, -10, -5, 0, +5, +10, and +15 degree with preprocessing of depth maps.

(a) Multi-view images of Break-dance

(b) Multi-view images of Home-Shopping

**Fig. 18. Multi-view generation with preprocessing of depth maps**

# VI. CONCLUSIONS

In this paper, we proposed a new 3-D natural video system using a layered representation of depth maps. After decomposing depth maps hierarchically, we reconstructed natural 3-D dynamic scenes based on grid, boundary, number

of layer (NOL), and feature point layers. Then, we rendered 3-D natural videos progressively according to consumers' capabilities. With the proposed system, we could render 3-D natural videos with reliable visual qualities in real time, but also compressed depth maps efficiently. We increased rendering speed for 3-D natural videos more than 30 times in comparison with a 3-D full modeling technique. Moreover, we reduced the amount of bits needed to code depth maps about 50 %, since we coded four layer images instead of original depth images. Finally, we could control visual qualities of 3-D natural videos adaptively according to multimedia platforms. In addition, we could generate multi-view images using an adaptive smoothing filter. We expect the proposed 3-D natural video system to be used for next-generation broadcasting or various 3-D applications.

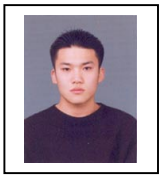
## ACKNOWLEDGMENT

This research was supported by MIC, Korea, under the ITRC support program supervised by IITA (IITA-2005-C1090-0502-0022) through the Realistic Broadcasting Research Center (RBRC) at the Gwangju Institute of Science and Technology (GIST).

## REFERENCES

[1] M. Reiner, "The role of haptics in immersive telecommunication environments," *IEEE Tran. Circuits Syst. Video Technol.*, vol. 14, no. 3, pp. 392-401, Mar. 2004
[2] W. Ijsselsteijn, H. De Rider, R. Hamberg, D. Bouwhuis, J. Freeman, "Perceived depth and the feeling of presence in 3DTV," *Displays*, vol. 18, pp. 207-214, 1998
[3] M. Kawashima, K. Yamamoto, and K. Kawashima, "Display and projection devices for HDTV," *IEEE Trans. Consumer Electron.*, vol. 34, no. 1, pp. 100-110, Feb. 1988
[4] A.. Redert, M. Op de Beeck, C. Fehn, W. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, I. Sexton, P. Surman, "ATTEST–Advanced Three-Dimensional Television System Technologies," *Proc. International Symposium on 3D Data Processing*, pp. 313–319, 2002
[5] K. Balasubramanian, "On the realization of constraint-free stereo television," *IEEE Trans. Consumer Electron.*, vol. 50, no. 3, pp. 895-902, Aug. 2004
[6] P. Bao and D. Gourlay, "Superview 3D image warping for visibility gap errors," *IEEE Trans. Consumer Electron.*, vol. 49, no. 1, pp. 177-182, Feb. 2003
[7] S.U. Yoon, S.Y. Kim, and Y.S. Ho, "A Framework for multi-view video coding using layered depth images," *Lecture Notes in Computer Science*, vol. 3767, pp. 431-442. 2005
[8] H. Shum and S. Kang, "A review of image-based rendering techniques," *Proc. Visual Communication and Image Processing,* pp. 2-13, 2000
[9] A. Ignatenko and A. Konushin, "A framework for depth image-based modeling and rendering," *Proc. Graphicon*, pp. 169-172, 2003
[10] S.M. Kim, J. Cha, J. Ryu, and K.H. Lee, "Depth video enhancement of haptic interaction using a smooth surface reconstruction," *IEICE Trans. on Information and System*, vol. E89-D, pp. 37-44, 2006.
[11] C. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 7, pp.675-684, 2000
[12] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *SIGGRAPH*, pp. 600-608., 2004
[13] S. Grewatsch and E. Muller, "Fast mesh-based coding of depth map sequences for efficient 3D video reproduction using OpenGL," *Proc. International Conference on Visualization, Imaging and Image Processing*, 2005.
[14] B. Chai, S. Sethuraman, H. Sawhey, and P. Hatract, "Depth map compression for real-time view-based rendering," *Pattern Recognition Letters*, vol. 25 pp. 755-766, 2004.
[15] C. Fehn, K. Schuur, P. Kauff and A. Smolic, "Coding results for EE4 in MPEG 3DAV," ISO/IEC JTC1/SC29/WG11 M9561, 2003.
[16] M. Deering, "Geometry compression," *SIGGRAPH*, pp. 13-20, 1995
[17] G. Taubin and J. Rossignac, "Geometry compression through topological surgery," *SIGGRAPH*, pp. 84-115. 1998
[18] T. Wiegand, M. Lightstone, D. Mukherjee, T.G. Campbell, and S.K. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard," *IEEE Trans. Circuit and System for Video Technology*, vol. 6, no. 9, pp. 182-190, 1996
[19] L. Zhang and W.J. Tam, "Stereoscopic image generation based on depth images for 3D TV," *IEEE Trans. Broadcasting*, vol. 51, pp. 191-199, 2005
[20] H. Hoppe, "Progressive meshes," *SIGGRAPH*, pp. 99-108, 1996
[21] J. Shade, S. Gotler, and R. Szeliski, "Layered depth images," *SIGGRAPH*, pp. 291-298, 1998
[22] M. Eck, T. DeRose, T. Duchamp, H. Hoppe, M. Lounsbery, ans W. Stuetzle, "Multi-resolution analysis of arbitrary meshes," *SIGGRAPH*, pp. 173-182, 1995
[23] J. Cha, S.M. Kim, S.Y. Kim, S. Kim, I. Oakley, J. Ryu, K.H. Lee, W. Woo, Y.S. Ho, "Client system for realistic broadcasting: a first prototype," *Lecture Notes in Computer Science*, vol. 3768, pp. 176-186, 2005
[24] 3DV systems, http://www.3dvssytems.com/, 2005
[25] C. Fehn, "Depth-image-based Rendering (DIBR), compression and transmission for a new approach on 3D TV," *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems*, vol. 5291, pp. 93-104, 2004

**Sung-Yeol Kim** received his B.S. degree in Information and Telecommunication engineering from Kangwon National University, Korea, in 2001 and M.S. degree in Information and Communication Engineering at the Gwangju Institute of Science and Technology (GIST), Korea, in 2003. He is currently working towards his Ph.D. degree in the Information and Communications Department at GIST, Korea. His research interests include digital signal processing, video coding, 3-D mesh representation, 3-D mesh compression, 3-D television, and realistic broadcasting.



**Sang-Beom Lee** received his B.S. degree in Electrical engineering from Kyungpook National University, Korea, in 2004. He is currently working towards his M.S. degree in the Information and Communications Department at GIST, Korea. His research interests include digital signal and image processing, video coding, 3-D mesh representation, 3-D television, and realistic broadcasting.

**Yo-Sung Ho** received both B.S. and M.S. degrees in electronic engineering from Seoul National University, Korea, in 1981 and 1983, respectively, and Ph.D. degree in Electrical and Computer Engineering from the University of California, Santa Barbara, in 1990. He joined the Electronics and Telecommunications Research Institute (ETRI), Korea, in 1983. From 1990 to 1993, he was with Philips Laboratories, Briarcliff Manor, New York, where he was involved in development of the advanced digital high-definition television (AD-HDTV) system. In 1993, he rejoined the technical staff of ETRI and was involved in development of the Korea direct broadcast satellite (DBS) digital television and high-definition television systems. Since 1995, he has been with the Gwangju Institute of Science and Technology (GIST), where he is currently a professor in the Information and Communications Department. His research interests include digital image and video coding, image analysis and image restoration, advanced coding techniques, digital video and audio broadcasting, 3-D television, and realistic broadcasting.