

# Multi-View Image Coding Using 3-D Voxel Models

Yongying Gao and Hayder Radha

Department of Electrical and Computer Engineering, Michigan State University

East Lansing, MI 48824, USA

Email: {gaoyongy, radha}@egr.msu.edu

**Abstract**—We propose a multi-view image coding system in 3-D space based on an improved volumetric 3-D reconstruction. Unlike existing multi-view image coding schemes, in which the 3-D scene information is represented by a mesh model as well as the texture data, we use a 3-D voxel model to represent the 3-D scene information of the images to be encoded. Furthermore, we propose important, yet simple, improvements to current 3-D voxel models; these improvements lead to significant coding gain within our 3-D voxel model based compression system. This system provides an elegant framework for employing powerful and mature coding techniques. In particular, we employ the H.264 and 3-D SPIHT coding methods for compressing the proposed 3-D voxel models. Our simulation results clearly illustrate the efficiency and potential of the proposed 3-D voxel model based system.

**Keywords**—multi-view image coding; 3-D voxel model; volumetric reconstruction

## I. INTRODUCTION

Multi-view image coding has been increasingly attracting attention for its crucial role in various applications, i.e., image-based rendering, medical volumetric data compression, and virtual reality. Studies in [1][2][3][4][4] show that multi-view image coding schemes using 3-D scene geometry information greatly improve the encoding efficiency, decoding speed and the rendering quality, compared with the conventional coding schemes employing only simple extension of 2-D compression.

However, there are still some aspects of the 3-D geometry-based coding schemes that can be improved. First, the mesh model and the texture data must be encoded separately. This requirement limits the flexibility of the coding scheme. Second, the generally used mesh model is suitable to represent 3-D objects of simple surface but difficult to represent objects of complicated surface, which are often shown in natural scenes. Third, the whole procedure of obtaining a mesh model is computationally complex [5][6][7][8].

We propose a multi-view coding system that operates directly in 3-D space and is based on volumetric 3-D reconstruction. The key difference of the proposed coding system from existing multi-view image coding schemes is that instead of representing the 3-D scene information of the images to be encoded by a mesh model as well as the texture data and encoding them, we use a 3-D voxel model to represent the 3-D scene information of the considered images and then encode the 3-D voxels directly. Furthermore, we propose important, yet simple, improvements to current approaches for volumetric reconstruction; these improvements lead to significant coding

gain within the proposed 3-D voxel model based coding system. Simulation results show the efficiency and potential of the proposed 3-D voxel model based system.

The remainder of this paper is organized as follows. In Section II, the framework for the proposed multi-view image coding system in 3-D space is introduced. Section III discusses our proposed volumetric 3-D reconstruction. Details on 3-D voxel model coding and residual coding are presented in Section IV and Section V. Section VI concludes this paper.

## II. FRAMEWORK FOR MULTI-VIEW IMAGE CODING IN 3-D SPACE

The framework of the proposed 3-D voxel model based multi-view coding system is shown in Fig. 1. The encoding part consists of:

1) *Volumetric 3-D Reconstruction*: We provide an algorithm for volumetric 3-D reconstruction, which is an improved version of Eisert's approach [5]. This step is one of the crucial steps in the system, since the quality of the 3-D voxel model impacts the efficiency of the 3-D data coding and the optional residual coding.

2) *3-D Model Encoding*: This step is another crucial step in the framework. We aim at encoding the 3-D scene voxel model (or simply the set of 3-D voxels) that represents the available multiple images in 3-D space.

3) *Coding of Camera Parameters*: This step is straightforward. The encoded data size of the camera intrinsic and extrinsic parameters is trivial compared with that of the 3-D coding and the residual coding.

4) *Residual coding*: This is an optional procedure but is anticipated to be required for high-quality applications. The residuals between the original images and re-projected images are computed and then encoded.

The decoding part is basically an inverse procedure of the encoding part, except that the corresponding block of the 3-D reconstruction in the encoding part is the re-projection process to recover the images from the decoded 3-D scene voxel model.

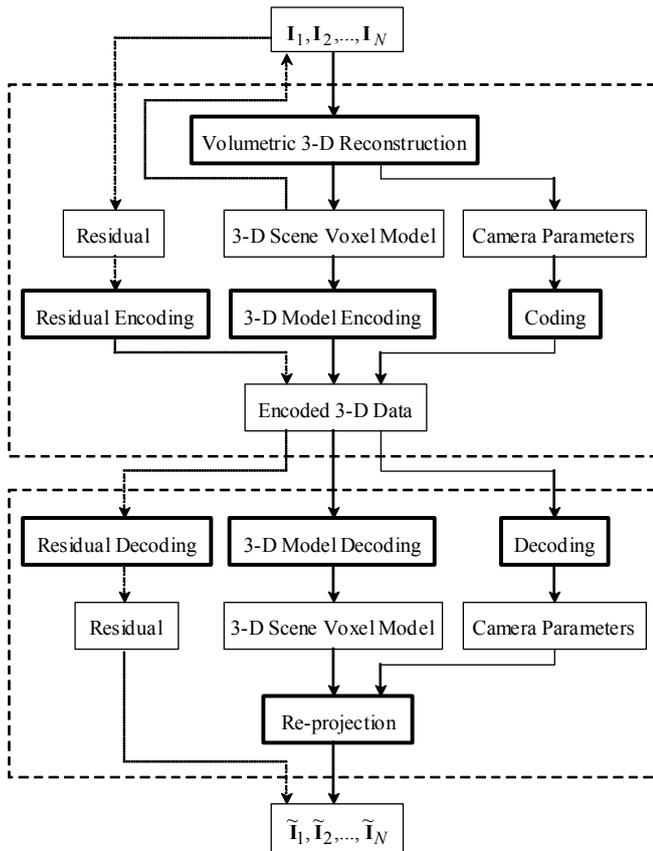


Fig. 1 The proposed framework for 3-D voxel model based multi-view image coding.

### III. IMPROVED VOLUMETRIC 3-D RECONSTRUCTION

We propose an algorithm for volumetric 3-D reconstruction from multiple calibrated images. Similar to Eisert’s approach [5] (referred as “the basic approach”), our approach proceeds in four successive steps: (1) volume initialization; (2) color hypothesis generation; (3) consistency check and hypothesis elimination; and (4) determination of the best color for the surface voxels. However, differing from the basic approach, we provide two improvements: (1) an enhanced hypothesis generation, and (2) a new measurement for pixel color difference.

#### A. Enhanced Hypothesis Generation

In step 2 of the basic approach, a voxel is determined to be “valid” for further consistency check if its associated hypotheses contain color values from at least two different images. However, many voxels that are not on the considered object surface are determined to be “valid” according to such a rule. To reduce the possibility of the “pseudo-valid” voxels, we increase the number of hypotheses of a “valid” voxel from 2 to  $K$  with  $K$  an integer larger than 2 but no larger than the maximum number of all available image pairs. Factors that may impact the value of  $K$  include the total number of the considered images, the camera parameters, the histogram of the considered images, and the resolution of the bounded 3-D space.

Another problem with the basic approach is that there is no processing in the consistency check for the occluded voxels. To solve this problem, in the re-projection of the 3-D voxel model, we remove those voxels in the 3-D model that are never assigned to image pixels.

#### B. A New Measurement for Color Difference

In step2 and 3 of the basic approach, the following inequality plays an important role:

$$\begin{aligned} &|r(u_j, v_j) - r(u_i, v_i)| + |g(u_j, v_j) - g(u_i, v_i)|, \\ &+ |b(u_j, v_j) - b(u_i, v_i)| < \Theta \end{aligned} \quad (1)$$

where  $(u_i, v_i)$  and  $(u_j, v_j)$  represent the coordinates of two pixels and  $\Theta$  is a pre-determined value.

Equation (1) provides a mechanism for measuring the difference of the color information between two pixels. However, the objective result using this measurement may not match the subjective result based on observations of human beings, since the RGB color system does not match physiological characteristics of the human visual system, i.e., the human eye is more sensitive to changes in brightness than to chromaticity changes. Therefore, we have modified the measurement for pixel color information difference based on the conversion from RGB to Y component (luminance):

$$\begin{aligned} &0.299 \times |r(u_j, v_j) - r(u_i, v_i)| + 0.587 \times |g(u_j, v_j) - g(u_i, v_i)| \\ &+ 0.114 \times |b(u_j, v_j) - b(u_i, v_i)| < \Theta \end{aligned} \quad (2)$$

#### C. Experimental Results

In Table 1, we compare three 3-D voxel models for a same test image sequence—the *cup*, which was also used in [5] and consists of 14 images with known camera calibration information. The first 3-D voxel model, named “VM3a”, was obtained using the basic approach<sup>1</sup>; the second one, named “VM5b”, was obtained using our first improvement; the third one, named “VM5c”, was obtained using both of our improvements.

TABLE 1 COMPARISON OF DATA SIZE AND THE AVERAGE PSNR<sup>2</sup> OF RECONSTRUCTED IMAGES AMONG THE OBTAINED THREE 3-D Voxel MODELS—VM3A, VM5B AND VM5C.

	VM3a	VM5b	VM5c
Voxel Number	146,005	86,064	82,622
Average PSNR of Rec. Images (dB)	16.73	19.48	20.26

<sup>1</sup> The VM3a was obtained using our software, in which the detailed steps may be different from those in [5].

<sup>2</sup> To better illustrate the reconstruction quality of the considered object, the PSNR is calculated within a bounding frame that just contains the considered object and neglects most part of the background. However, compared with the PSNR that is calculated over the whole image, the presented PSNR shows a lower value.

Table 1 shows that both the quality of reconstructed images and the data size of the 3-D voxel model are significantly improved by employing the proposed two improvements.

We also display in Fig. 2 the original images 1, 2 and 3 (from the left to the right) in the *cup* sequence and corresponding reconstructed images from re-projection of the obtained three 3-D voxel models.



Fig. 2 1<sup>st</sup> row: selected original images; 2<sup>nd</sup> row: reconstructed images from VM3a; 3<sup>rd</sup> row: reconstructed images from VM5b; 4<sup>th</sup> row: reconstructed images from VM5c.

Fig. 2 clearly shows that the quality of the reconstructed images is gradually improved from VM3a to VM5c. VM5c shows the best performance among the three models. We also observed that for all the three voxel models, the reconstruction quality of image 1 and 2 is better than that of image 3. This observation is not surprising because image 1 and 2 belong to a group of “dense” camera viewing positions, which refers to a large number of camera viewing positions gathering within certain spatial volume, while image 3 belongs to a group of “sparse” camera viewing positions.

#### IV. 3-D VOXEL MODEL CODING

Having obtained the 3-D voxel model, we target encoding the 3-D scene voxel model that represents in 3-D space all the available images. We employ two approaches: (1) 3-D wavelet-based SPIHT coding scheme [9][10]; and (2) H.264 video coding standard based coding scheme, considering the 3-D voxel model as a set of highly correlated “video frames”.

In our 3-D voxel model, many voxels within the predefined volume are not on the 3-D object surface and can be marked as “useless”. We label the “useful” voxels and the label set is stored and transmitted along with the 3-D voxel model as side information. All the “useless” voxels are assigned the average color value of all the “useful” voxels to reduce the high-frequency energy.

We provide some simulation results of 3-D voxel model coding here. To simplify the problem, we consider only the luminance information of the test image sequence “the *cup*” (which was also used in [5]) and the obtained three 3-D voxel models.

Fig. 3 depicts the coding performance of the three 3-D voxel models using the 3-D wavelet-based SPIHT coding scheme. The shown bit rate includes both the encoded label data and the encoded pre-processed 3-D data. For a certain 3-D scene model, the size of encoded label data is fixed.

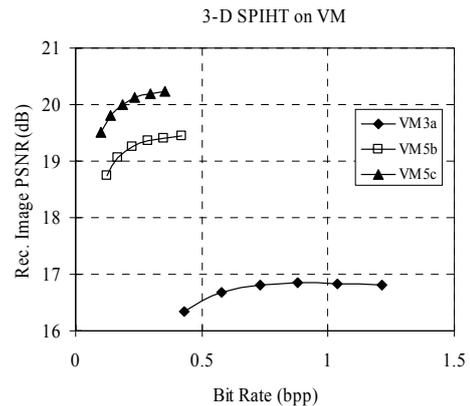


Fig. 3 Rate-PSNR curves of the 3-D wavelet-based SPIHT coding for the three 3-D voxel models. The x-axis represents the bit rate of the encoded 3-D voxel model; the y-axis represents the average image quality over the available reconstructed images from re-projection of the decoded 3-D voxel model.

We can conclude from Fig. 3 that the coding performance of the VM5c is the best among the three models. Similar conclusions can be made from the coding performance of the three 3-D voxel models using H.264 based coding scheme.

Next, we compared the coding performance for the same 3-D voxel model using the two proposed coding schemes, as shown in Fig. 4.

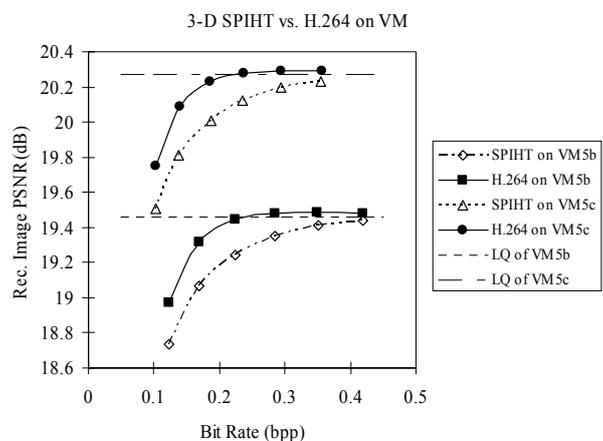


Fig. 4 Comparison of the rate-PSNR 3-D voxel curves between the 3-D SPIHT and the H.264-based coding scheme.

There are two observations from Fig. 4. First, the H.264-based coding scheme outperforms the 3-D SPIHT coding scheme for both of the 3-D voxel models. Second, for the same 3-D voxel model, both of the two rate-PSNR curves approach the “Lossless Quality (LQ)” (19.46 dB for VM5b and 20.27 dB for VM5c). “LQ” represents the quality of the reconstructed images directly from re-projection of the 3-D voxel model without encoding and decoding. However, the rate-PSNR curve of the H.264-based coding scheme converges to the LQ more quickly than that of the 3-D SPIHT coding scheme does.

## V. 3-D VOXEL MODEL AND RESIDUAL CODING

In the proposed coding system, the residual coding will be required for high-quality reconstruction of original images in many applications. However, the residual between the original images and re-projected images is rather different from residual signals in other coding systems. In detail, the origin of the residual is the difference between the true 3-D scene structure and the estimated 3-D voxel model. One voxel that contains incorrect color information will lead to correlated errors among all the considered images. Hence, the residual images show correlations with each other. To de-correlate the residual images, we also employ the H.264 video coding standard or the 3-D SPIHT coding scheme to the residual images.

Fig. 5 provides simulation results for the coding performance of the residual de-correlation and regulation. For each considered 3-D model, we employed the 3-D SPIHT based coding scheme and the H.264 based coding scheme for both the 3-D voxel model coding and the residual coding. For each rate-PSNR curve, we chose a fixed bit rate for the 3-D model coding.

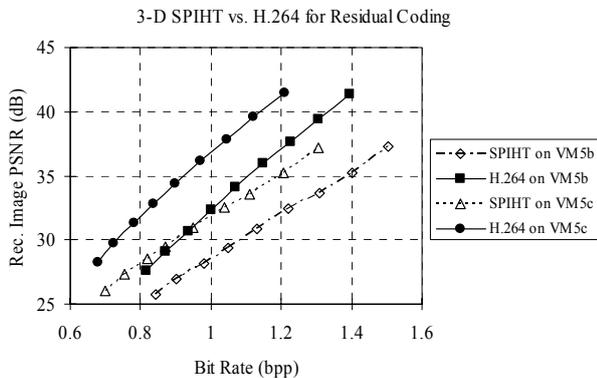


Fig. 5 Comparison of the rate-PSNR curves between the 3-D SPIHT residual coding and the H.264-based residual coding for VM5b and VM5c.

Fig. 5 shows that the performance of the H.264-based residual coding is better than that of the 3-D SPIHT residual coding for the considered two 3-D voxel models. Moreover, for the same 3-D voxel model, the improved coding efficiency of the H.264-based coding scheme over the 3-D SPIHT coding scheme becomes greater and greater with the increase of the bit rate.

## VI. CONCLUSIONS

In this paper, we proposed a multi-view image coding system in 3-D space based on an improved volumetric 3-D reconstruction. Unlike existing multi-view image coding schemes, in which the 3-D scene information of the images to be encoded is represented by a mesh model as well as the texture data, we utilize a 3-D voxel model to represent the 3-D scene information of the considered images and then encode the 3-D voxel model. We employed the H.264 video coding standard and 3-D SPIHT coding method for compressing three proposed 3-D voxel models. Our simulation results clearly illustrate the efficiency and potential of the proposed 3-D voxel model based system.

## REFERENCES

- [1] S. M. Seitz and C. M. Dyer, “View morphing”, *Proc. ACM Conf. Computer Graphics’96*, pp. 21-30, 1996.
- [2] D. Wood, D. Azuma, K. Aldinger, B. Curless, T. Duchamp, D. Salesin and W. Stuetzle, “Surface light fields for 3D photography”, *Proc. of ACM Conf. Computer Graphics’00*, pp. 287-296, 2000.
- [3] H. Schirmacher, W. Heidrich and H.-P. Seidel, “High-quality interactive lumigraph rendering through warping”, *Proc. of Graphics Interface2000*, pp. 87-94, 2000.
- [4] M. Magnor, P. Ramanathan and B. Girod, “Multi-view coding for image-based rendering using 3-D scene geometry”, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 11, pp. 1092-1106, 2003.
- [5] P. Eisert, E. Steinbach and B. Girod, “Multi-hypothesis, volumetric reconstruction of 3-D objects from multiple calibrated views”, *Proc. of IEEE Conf. on Acoustics, Speech and Signal Processing’1999*, pp. 3509-3512, 1999.
- [6] W. E. Lorensen and H. E. Cline, “Marching cubes: a high resolution 3D surface construction algorithm”, *Proc. of ACM Conf. Computer Graphics’87*, pp. 163-169, 1987.
- [7] H. Hoppe, “Progressive meshes”, *Proc. of ACM Conf. Computer Graphics’96*, pp. 99-108, 1996.
- [8] M. Magnor and B. Girod, “Fully embedded coding of triangle meshes”, *Proc. of Vision, Modeling and Visualization’1999*, pp. 253-259, 1999.
- [9] B.-J Kim, Z. Xiong and W. A. Pearlman, “Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT)”, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, no. 8, pp. 1374-1387, 2000.
- [10] Z. Xiong, X. Wu, S. Cheng and J. Hua, “Lossy-to lossless compression of medical volumetric data using three-dimensional integer wavelet transforms”, *IEEE Trans. on Medical Imaging*, vol. 22, no. 3, pp. 459-470, 2003.