

Multi-View Image Coding in 3-D Space Based on 3-D Reconstruction

Yongying Gao and Hayder Radha

Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48823
email: {gaoyongy, radha}@egr.msu.edu

Abstract— In this paper, we propose a multi-view image coding system that operates directly in 3-D space, and it is based on 3-D scene 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 use a 3-D voxel model to represent the 3-D scene information of the considered images and then encode the 3-D voxel model. There are several advantages of the 3-D voxel model over the mesh model as well as the texture data. Experimental results show the potential of our proposed multi-view image coding system. 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.

Keywords — multi-view image coding, volumetric 3-D reconstruction, 3-D data coding

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. These applications have the common goal of handling a large number of highly correlated 2-D images. Studies in [1][2][3][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 scene geometry information and the image data must be encoded separately. This requirement limits the flexibility of the coding scheme, since the decoding of the 3-D geometry information must be completed prior to the decoding of the image (texture) data. Second, the generally used 3-D geometry representations, such as the mesh model used in

existing 3-D geometry-based multi-view coding schemes, are 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 3-D geometry information is computationally complex [5][6][7][8].

The existing problems stated above have motivated our study in further combining the 3-D scene reconstruction into multi-view coding. We propose a multi-view coding system that operates directly in 3-D space and is based on 3-D scene reconstruction. The key difference of our proposed multi-view coding system from existing multi-view coding schemes is that instead of representing the 3-D scene information of the consider images by the mesh model as well as the texture data and encoding the 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 and then encode the 3-D voxel model. Simulation results show a clear potential for the proposed 3-D voxel model based coding approach.

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 are presented in Section IV. In Section V, we discuss the residual coding. Section VI concludes this paper.

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

The framework of the proposed 3-D space multi-view coding system is shown in Figure 1.

The encoding part consists of three necessary blocks and one optional block:

1) *Volumetric 3-D reconstruction*: The input to the encoding part is a set of images, denoted by $\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_N$. The N images are then fed into the block of volumetric 3-D reconstruction to obtain the 3-D scene voxel model. 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 our multi-view coding system, since the quality of the reconstructed 3-D voxel

model significantly impacts the coding efficiency of the 3-D data coding and the optional residual coding.

2) *3-D Model Encoding*: This step is another crucial step in our proposed coding framework. Generally speaking, we aim at encoding the 3-D scene model that represents the available multiple images in 3-D space. We propose two possible approaches: (1) employing the H.264 video coding standard for compressing the 3-D voxel; and (2) 3-D wavelet-based SPIHT coding scheme.

3) *Coding of Camera Parameters*: The obtained camera intrinsic and extrinsic parameters are quantized for encoding purposes. This step is straightforward and less important than 1), 2) and 4), because the encoded data size of the camera parameters is trivial compared with the encoded data size of the 3-D coding and the residual coding.

4) *Residual coding*: This is an optional procedure in our system. However, we anticipate that the residual coding will be required for high-quality applications. In the case that the quality of reconstructed images from re-projection of the 3-D scene model does not meet the requirement of the specific application, the residuals between the original images and re-projected images are computed and then encoded.

The final encoded 3-D data includes the encoded 3-D data, the encoded camera parameters and the (optional) encoded residual data.

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. The target of re-projection is to recover the images from the decoded 3-D scene voxel model and the camera parameters.

III. IMPROVED VOLUMETRIC 3-D RECONSTRUCTION

We propose an algorithm for volumetric 3-D reconstruction from multiple calibrated images. In this approach, all operations are performed on voxels, which are the basic elements of the 3-D object. Therefore, we avoid the search for corresponding pixels between two images and the fusion of incomplete depth estimates.

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 based on physiological characteristics of the human visual system.

Details on the two improvements are beyond the scope of this paper. Here we provide some experimental results of the improved volumetric 3-D reconstruction. 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; 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 PSNR¹ 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

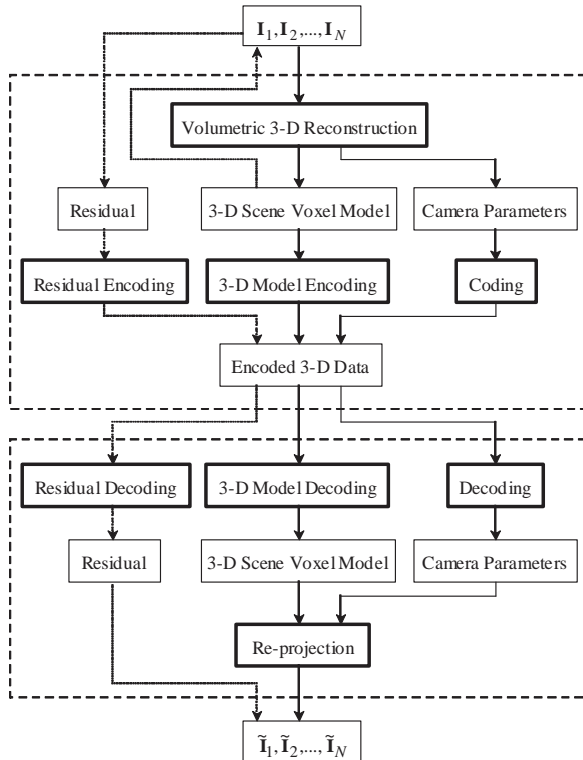


Figure 1 The proposed framework for multi-view image coding in 3-D space.

¹ The calculation of PSNR is performed within a bounding frame that just contains the considered object and neglects most part of the background.

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 Figure 2 the original images 3 and 6 in the *cup* sequence and corresponding reconstructed images from re-projection of the obtained three 3-D voxel models.

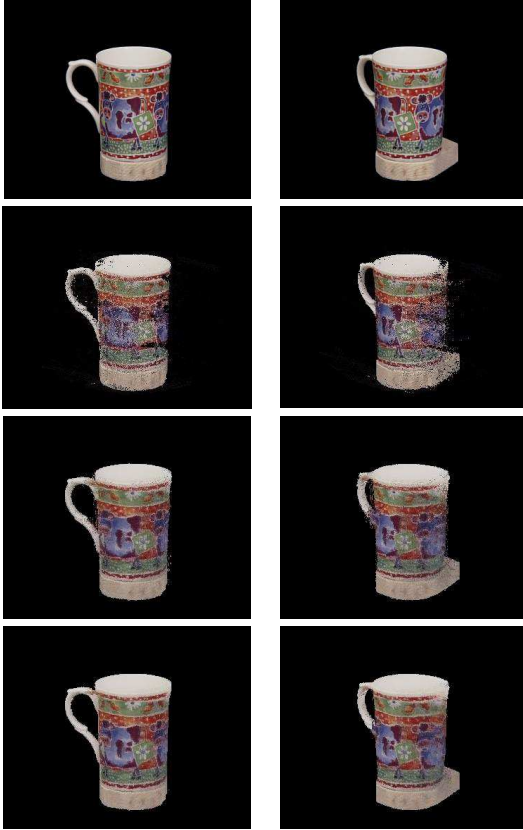


Figure 2 1st row: selected original images; 2nd row: reconstructed images from VM3a; 3rd row: reconstructed images from VM5b; 4th row: reconstructed images from VM5c.

Figure 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.

IV. 3-D VOXEL MODEL CODING

Having obtained the 3-D voxel model by volumetric 3-D reconstruction, we target encoding the 3-D scene voxel model that represents in 3-D space the available multiple images. We propose two possible approaches: (1) 3-D wavelet-based SPIHT coding scheme; and (2) H.264 based coding scheme. The idea behind the proposed approaches (1) and (2) is that a video stream sequence, volumetric data (e.g., a set of medical

images), and the proposed 3-D voxel model are common in (a) they all can be regarded as three dimensional data, and (b) correlations exist along all the three dimensions. Currently, we have obtained experimental results using the H.264-based and 3-D SPIHT-based coding schemes.

A. Label Coding for 3-D Voxel Model

Before we start encoding the 3-D voxel model, it is necessary to consider the characteristics of the 3-D voxel model for any needed modification of the algorithm that we will apply. For example, in common volumetric data, every element contained in the volume is useful for the purpose of representation. On the contrary, in our 3-D voxel model, a vast majority of the voxels within the predefined volume is not on the 3-D object surface and can be marked as “useless” in representing the considered object. Hence, we must find a way to identify the “useful” and “useless” voxels in encoding the 3-D voxel model.

To solve the above problem, we label all of the “useful” voxels and the label set is stored and transmitted along with the 3-D voxel model as side information. With the label data, we assign all the “useless” voxels the average color value of all the “useful” voxels to reduce the high-frequency energy. The pre-processed 3-D voxel model is then applied to the 3-D coding scheme.

B. 3-D Wavelet-based SPIHT Coding Scheme

The well-known SPIHT image coding algorithm [9] is among state-of-the-art image coding techniques. Kim et al. [10] applied a 3-D extension of the SPIHT for a low bit rate embedded video coding scheme. The 3-D SPIHT was also successfully employed in medical volumetric data compression [11]. We employed the 3-D wavelet-based SPIHT coding scheme to our obtained 3-D voxel models.

C. H.264-based 3-D Data Coding Scheme

The 3-D voxel models generated by our multi-view image coding system can be considered as a set of highly correlated “video frames”. Therefore, the motion compensation that is commonly applied in video coding schemes helps to exploit the correlations along the z -dimension of our 3-D voxel model. In our simulations, we applied the H.264 video coding standard [12] to our obtained 3-D voxel models.

D. Simulation Results

We provide a set of simulation results of 3-D voxel model coding. To simplify the problem, we consider only the luminance information of the original image sequence “the *cup*” and the obtained three 3-D voxel models—VM3a, VM5b and VM5c. Focusing on the luminance performance is both reasonable and widely acceptable because of the sensitivity of human eyes to changes in luminance.

Figure 3 and Figure 4 depict the coding performance for the three 3-D voxel models using 3-D wavelet-based SPIHT coding scheme and H.264-based coding scheme, respectively. In the SPIHT coding scheme, the number of frames in one segment is 16. 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 and does not impact the shape of the rate-PSNR curve of the used coding scheme.

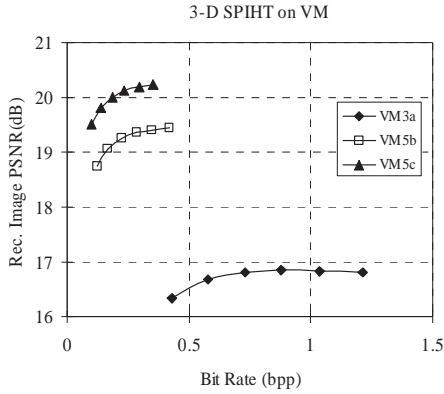


Figure 3 Rate-PSNR curves of the 3-D wavelet-based SPIHT coding for the three 3-D voxel models. The x-axis represents the bit rate of the encoded 3-D voxel model; the y-axis represents the average image quality over the available reconstructed images from re-projection of the decoded 3-D voxel model.

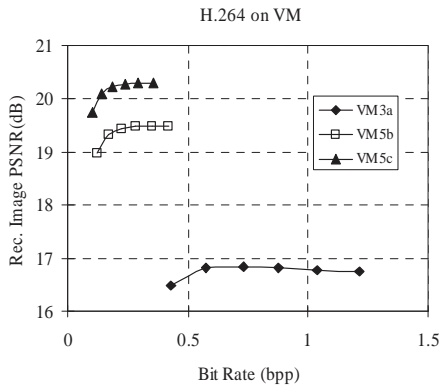


Figure 4 Rate-PSNR curves of the H.264-based coding for the three 3-D voxel models.

We can conclude from Figure 3 and Figure 4 that the coding performance for the VM5c is the best among the three models, regardless if the 3-D SPIHT coding scheme or the H.264-based coding scheme is employed. This observation is consistent with the experimental results shown in Table 1 that the VM5c performs the best among the three 3-D voxel models according to both the model size and the quality of reconstructed

images directly obtained from the re-projection process. In particular, the coding performance for the VM3a is significantly worse than that for the VM5b and the VM5c. This observation is reasonable since the performance for the VM5b and VM5c is close to each other while the performance for the VM3a is far from them.

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

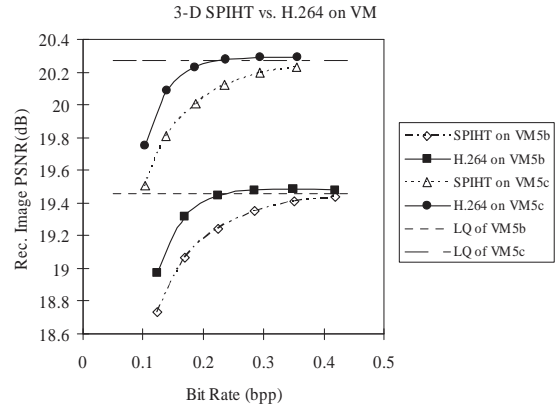


Figure 5 Comparison of the rate-PSNR curves between the 3-D SPIHT and the H.264-based coding scheme for VM5b and VM5c. The x-axis represents the bit rate of the encoded 3-D voxel model; the y-axis represents the average image quality over the available reconstructed images from re-projection of the decoded 3-D voxel model. The “LQ” is the abbreviation of “Lossless Quality”, which represents the quality of the reconstructed images directly from re-projection of the corresponding 3-D voxel model without encoding and decoding processes.

There are two observations from Figure 5. 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 (19.46 dB for VM5b and 20.27 dB for VM5c, shown in Figure 5-9). However, the rate-PSNR curve for the H.264-based coding scheme converges to the Lossless Quality more quickly than that for the 3-D SPIHT coding scheme does.

V. RESIDUAL CODING

In the proposed 3-D space multi-view coding system, the residual coding will be required for high-quality reconstruction of original images in many applications, since the image reconstruction directly from re-projection of the 3-D scene model is far from perfect. However, in our multi-view coding system, the residual between the original images and re-projected images is

quite different from that in video coding schemes in two aspects. These two aspects require special considerations in coding the residual data.

A. Residual De-correlation

In many video coding schemes, the residual data shows little correlation among neighboring frames. In our case, 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 in our multi-view coding scheme show correlations with each other. To de-correlate the residual images, we propose to employ the H.264 video coding standard or 3-D SPIHT coding scheme to the residual images

B. Residual Regulation

Another character of the residual data in our case is that it can be distributed in a larger range of values, unlike the residual data in many video coding schemes, which usually has a smaller variation. For instance, in the widely used 8-bit representation of basic color component (either YUV or RGB color system), the valid value is between 0 and 255. However, the residual data between the original images and the re-projected images can be as least as -255 or as great as 255 . To resolve this issue when using coding standards (e.g., H.264) that operates on 8-bit pixels, we regulate the residual data by shifting and rounding-off, named “*residual rescaling*”:

$$R_r = \lfloor (R + 255) / 2 \rfloor, \quad (1)$$

where R and R_r represent the original residual and the rescaled residual, respectively. Now the rescaled residual is located in $[0, 255]$ and can be represented by 8 bits. The final reconstructed image is calculated by:

$$\hat{I} = I_{rep} + (2R_r - 255), \quad (2)$$

where \hat{I} represents the final reconstructed image, I_{rep} represents the re-projected images from the 3-D voxel model, and R_r represents the rescaled residual image.

The residual rescaling does not increase the size of residual data at the expense of ignoring the least significant bit to rescale the residual from 9-bit representation to 8-bit representation and resulting in a “lossy” residual data.

C. Simulation Results

Figure 6 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.

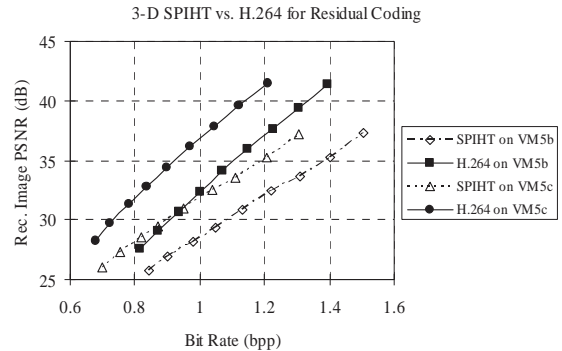


Figure 6 Comparison of the rate-PSNR curves between the 3-D SPIHT residual coding and the H.264-based residual coding for VM5b and VM5c, using the residual rescaling technique. The x-axis represents the bit rate of the encoded 3-D data and the encoded residual data; the y-axis represents the average image quality over the available final reconstructed images from the re-projected images (from the decoded 3-D voxel model) plus the compensation (from the decoded residual data).

Figure 6 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 and discussed in detail two crucial functional blocks of it: 3-D data coding and residual coding.

Unlike existing multi-view image coding schemes, in which the 3-D scene information of the images to be encoded is represented by the mesh model as well as the texture data, we adopt a 3-D voxel model to represent the 3-D scene information of the considered images and then encode the 3-D voxel model for the purpose of storage and transmission. There are several advantages of the 3-D voxel model. First, the 3-D voxel model is much simpler than the mesh model in structure. Second, recovering the original images or generating synthetic images from the 3-D voxel model is straightforward by the re-projection of the 3-D model; meanwhile image reconstruction from the mesh model requires mapping the texture data to the mesh model. Third, since the 3-D voxel model is an extension from 2-D data to 3-D data, many existing techniques for the image/video coding can be applied for the coding of the 3-D voxel model. We have

employed the H.264 coding standard and the 3-D SPIHT coding scheme in our experiments. Experimental results show the potential of our proposed multi-view image coding system. Furthermore, we proposed 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, as shown in the simulation results.

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] A. Said and W. A. Pearlman, "A new, fast and efficient image codec based on set partitioning in hierarchical trees", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 6, no. 3, 1996.
- [10] 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.
- [11] 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.
- [12] Joint Video Team of ITU-T and ISO/IEC, "Draft ITU-T recommendation and final draft international standard of joint video specification", *ITU-T Recommendation H.264 - ISO/IEC 14496-10 AVC*, March 2003.