

Establishing the Scene Voxel Model of 3-D Objects from Multiple Calibrated Images

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 — We proposed an algorithm for establishing a 3-D scene voxel model of 3-D objects from multiple calibrated images. Differing from recent approaches for volumetric 3-D reconstruction, we made two modifications to improve the performance: first, we proposed an enhanced hypothesis generation and consistency check; second, we used a new measurement for pixel color information difference based on the physiological characteristics of the human visual system. Experimental results show that both the data size of final 3-D model and the quality of reconstructed images are significantly improved by employing the proposed two improvements. In addition, we discuss the influence of camera intrinsic parameters on the volumetric 3-D reconstruction.

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

I. INTRODUCTION

For many decades, capturing and display of visual data have been confined to 2-D techniques and methods. Moving from 2-D to 3-D visual representation is naturally a challenging task, and hence, current and emerging 3-D visual applications rely (in a significant way) on well-established 2-D visual sources and related methods. One widely used approach is to reconstruct the 3-D scene from multiple images. The recovered 3-D scene information can be used in many applications, such as virtual reality and multimedia applications.

Eisert et al. [1][2] proposed a multi-hypothesis volumetric reconstruction to obtain the 3-D scene geometry. In this approach, a volumetric model of the considered 3-D object is established by a multi-hypothesis testing of the re-projection of the object surface voxels with the available camera views. All operations are performed on voxels, and not on image pixels. Therefore, there is no need for the search for corresponding points and the fusion of incomplete

depth estimates, which are usually required in obtaining a 3-D surface model from depth maps [3][4].

In this paper, we propose an algorithm for establishing a 3-D scene voxel model to represent in 3-D space all the available multiple images. Similar to Eisert's approach, our approach is based on the multi-hypothesis check of the re-projection of the object surface voxels back to the image planes. However, differing from Eisert's algorithm, we present (1) an enhanced hypothesis generation and consistency check and (2) a new measurement for pixel color information difference based on the physiological characteristics of the human visual system. Experimental results show that both the quality of reconstructed images and the data size of final 3-D model are significantly improved by employing the proposed two improvements. In addition, we discuss the influence of camera intrinsic parameters on the volumetric 3-D reconstruction.

The remainder of this paper is organized as follows. Section II provides details on the proposed improvements for the volumetric 3-D reconstruction. Experimental results are shown in Section III to compare the obtained 3-D voxel models and reconstructed images between the basic approach for the volumetric 3-D reconstruction and the advanced approach with the proposed improvements. Section IV concludes this paper.

II. ESTABLISHING THE 3-D SCENE VOXEL MODEL FROM MULTIPLE CALIBRATED IMAGES

Three assumptions associated to the 3-D scene voxel model (which are not mentioned in Eisert's approach) are made for the proposed volumetric 3-D reconstruction. First, we assume that the considered images were captured under the perfect perspective projection of a pinhole camera. This assumption indicates that the considered images contain no aberrations caused by optical effects, such as radial distortion, spherical aberration or chromatic aberration. Second, we assume that the light condition is the same around the considered 3-D scene/object. Third, we assume that the considered 3-D scene/object is made

from materials of constant refractive index and isotropic reflection property. These two assumptions indicate that the luminance and chrominance of the same part of the 3-D scene displayed in different images were not impacted by the different camera viewing positions.

Similar to Eisert's approach, our approach proceeds in four successive steps:

- 1) volume initialization;
- 2) color hypothesis generation for all voxels from all available images;
- 3) consistency check and hypothesis elimination considering all the images;
- 4) determination of the surface voxel color.

Details of the four steps can be found in [2]. In this paper, we discuss only the proposed modifications in step 2) and 3).

A. Enhanced Hypothesis Generation and Consistency Check

Based on the discretization of the defined volume in step 1), the projection from a 3-D point $[x, y, z]^T$ to a pixel $[u_i, v_i]^T$ in the i -th view ($i=1,2,\dots,N$, with N the number of all available images) is expressed as

$$s_i \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} [\mathbf{R}_i \mathbf{t}_i] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \quad (1)$$

where α_u and α_v represent the focal lengths in pixels along the vertical and horizontal direction respectively, (u_0, v_0) are the coordinates of the principal point, and \mathbf{R}_i and \mathbf{t}_i represent the rotation matrix and translation vector of the i -th view, respectively. Thus, we obtain

$$\begin{cases} u_i = \alpha_u \frac{r_{11}x + r_{12}y + r_{13}z + t_x}{r_{31}x + r_{32}y + r_{33}z + t_z} + u_0 \\ v_i = \alpha_v \frac{r_{21}x + r_{22}y + r_{23}z + t_y}{r_{31}x + r_{32}y + r_{33}z + t_z} + v_0 \end{cases}, \quad (2)$$

where $\mathbf{R}_i = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$ and $\mathbf{t}_i = [t_x, t_y, t_z]^T$.

A hypothesis H_{lmn}^k for a voxel V_{lmn} is

$$H_{lmn}^k = [r(u_i, v_i), g(u_i, v_i), b(u_i, v_i)], \quad (3)$$

where $[u_i, v_i]^T$ represents the pixel position of the projection of the voxel center $[x_{l0}, y_{m0}, z_{n0}]^T$ into the i -th image view ($i \in [1, 2, \dots, N]$), and $r(\bullet)$, $g(\bullet)$ and $b(\bullet)$ represent the red, green and blue color

components. The relationship between $[u_i, v_i]^T$ and $[x_{l0}, y_{m0}, z_{n0}]^T$ is shown in Eqn. (1) and (2).

According to [1], the hypothesis H_{lmn}^k is associated with voxel V_{lmn} if the projection of V_{lmn} into at least one other camera view $j \neq i, j=1, 2, \dots, N$ leads to an absolute color difference that is less than a predefined threshold Θ

$$\begin{aligned} & \left| r(u_j, v_j) - r(u_i, v_i) \right| + \left| g(u_j, v_j) - g(u_i, v_i) \right| \\ & + \left| b(u_j, v_j) - b(u_i, v_i) \right| < \Theta \end{aligned} \quad (4)$$

For each view $i, i=1, 2, \dots, N$, the hypothesis H_{lmn}^k that satisfies Eqn. (4) is stored with the color taken from view i according to Eqn. (3).

At this stage, we have no knowledge of the 3-D object geometry and cannot determine whether a voxel is visible or not. We therefore need to remove those hypotheses that do not correspond to the correct color of the considered object surface by consistency check and hypothesis elimination.

In the current algorithm [1], a voxel is determined to be "valid" for further consistency check if its associated hypotheses contain the 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. Unfortunately, experiments show that not all these "pseudo-valid" voxels can be detected in the consistency-check step, especially those that are outside the object surface. 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 is an integer larger than 2 but no larger than the maximum number of all available image pairs. Factors that may impact the selection of 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. In our approach, we choose $K=3$. By using this method, we can remove a significant number of the voxels that are outside the considered object.

Another problem with the current approach [1] is that there is no special processing in the consistency check for the occluded voxels; that is, voxels that are inside the considered object. This problem does not impact the quality of the reconstructed images but makes the 3-D voxel model contain a considerable number of useless voxels. This impact is undesirable for our final goal of data compression. To solve this problem, in the re-projection of the 3-D voxel model back to the image planes, we remove those voxels that are never assigned to image pixels. By using this method, we can remove a large number of occluded voxels from the 3-D model.

B. A New Measurement for Pixel Color Information Difference

Eqn. (4) plays an important role in hypothesis generation and consistency check. It provides a mechanism for measuring the difference of the color information between two pixels. However, the objective result of pixel color information difference 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. This character of the human visual system has also led to the YUV (one luminance and two chrominance components) color image format in many of the standardized video coding schemes.

Therefore, we have modified the measurement for pixel color information difference based on the Y component (luminance). The conversion from RGB to Y is given as below:

$$Y = 0.299 \times r + 0.587 \times g + 0.114 \times b. \quad (5)$$

Eqn. (5) shows that the three components r, g, and b contribute quite different to the luminance, i.e., the green component impacts the luminance the most (it is why we say that the human eyes are more sensitive to green color than others.). Hence, we modified Eqn. (4) to a weighted summation of the absolute difference of the r, g, and b between two pixels, as shown below:

$$\begin{aligned} & 0.299 \times |r(u_j, v_j) - r(u_i, v_i)| \\ & + 0.587 \times |g(u_j, v_j) - g(u_i, v_i)|, \quad (6) \\ & + 0.114 \times |b(u_j, v_j) - b(u_i, v_i)| < \Theta \end{aligned}$$

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

C. Influence of the Camera Intrinsic Parameters on Volumetric 3-D Reconstruction

This section discusses the influence of inaccurate camera intrinsic parameters on the 3-D voxel model and image reconstruction. Similarly, the problem of influence of the position of the principal point on 3-D reconstruction was discussed in [5].

We suppose that the camera calibration matrix

$$\begin{bmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \text{ in Eqn. (1) is taken place by } \begin{bmatrix} \alpha_u + \Delta\alpha_u & 0 & u_0 + \Delta u_0 \\ 0 & \alpha_v + \Delta\alpha_v & v_0 + \Delta v_0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ then for the same}$$

3-D point $[x, y, z]^T$, the coordinates of the projected image pixel become

$$\begin{cases} u'_i = (\alpha_u + \Delta\alpha_u) \frac{r_{11}x + r_{12}y + r_{13}z + t_x}{r_{31}x + r_{32}y + r_{33}z + t_z} + (u_0 + \Delta u_0) \\ v'_i = (\alpha_v + \Delta\alpha_v) \frac{r_{21}x + r_{22}y + r_{23}z + t_y}{r_{31}x + r_{32}y + r_{33}z + t_z} + (v_0 + \Delta v_0) \end{cases} \quad (7)$$

The difference between the correct and inaccurate coordinated of the projected image pixel is shown as below:

$$\begin{cases} \Delta u_i = u'_i - u_i = \Delta\alpha_u \frac{r_{11}x + r_{12}y + r_{13}z + t_x}{r_{31}x + r_{32}y + r_{33}z + t_z} + \Delta u_0 \\ \Delta v_i = v'_i - v_i = \Delta\alpha_v \frac{r_{21}x + r_{22}y + r_{23}z + t_y}{r_{31}x + r_{32}y + r_{33}z + t_z} + \Delta v_0 \end{cases} \quad (8)$$

We make two conclusions from Eqn. (8): first, the shifting of the principal point from its true position causes a uniform shifting of the projected image pixels, no matter what the coordinates of the projecting 3-D point are. Hence, theoretically, the errors in the position of the principal point will not introduce any distortion of the obtained 3-D voxel model. Consequently, the reconstructed images will be the same as those from the 3-D voxel model that is obtained from the true intrinsic parameters; second, the errors of the coordinates of the projected image pixel caused by the inaccurate focal lengths are determined by not only the errors in the focal lengths, but the camera relative motion as well as the coordinates of the projecting 3-D point. Therefore, the errors in the focal lengths will significantly degenerate the volumetric 3-D reconstruction and result in a poor 3-D voxel model.

III. EXPERIMENTAL RESULTS

We provide experimental results for the algorithm for volumetric 3-D reconstruction described in Section 2. The test image sequence, known as the *cup* sequence, was downloaded from [6]. It consists of a total of 14 images (288×352) with known camera calibration information (camera intrinsic parameters, camera relative motion). The original images 3, 6, 11, 14 are shown in Figure 1.

The first voxel model, named “VM3a”, was obtained using the basic approach [1] without our modification¹. The voxel resolution for the initial volume is chosen as 160×160×160. VM3a contains 146,005 voxels. We re-projected the VM3a back to image planes for the same camera viewing positions of all the original

¹ The experimental results of the basic approach are obtained using our software.

images. The average *Peak Signal-to-Noise-Ratio* (PSNR) of the reconstructed 14 images is 16.73 dB (the calculation of PSNR is performed within a bounding frame that just contains the considered object and neglects most part of the background). We show in Figure 2 the reconstructed images corresponding to the images in Figure 1.



Figure 1 The original images 3, 6, 11, and 14 of the cup sequence

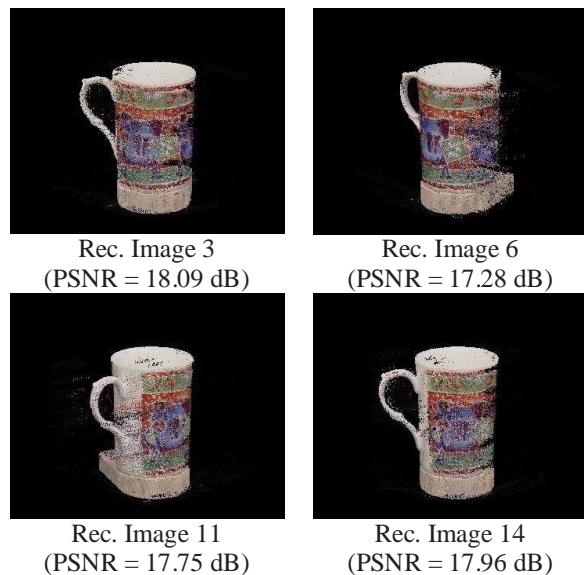


Figure 2 The reconstructed images resulting from re-projecting the VM3a back to image planes for the same camera viewing positions as in Figure 1, as well as the value of PSNR for each reconstructed image.

The re-projection of the 3-D voxel model is accomplished as follows. First we re-project each voxel to all the available images. Then for each pixel in each

image that is a result of the re-projection of one or more voxels, we assign it the color value of the associated voxel that has the smallest depth to the considered image plane.

The experimental results for volumetric 3-D reconstruction in Figure 2 show that the obtained 3-D voxel model as well as the known camera calibration information can represent and reconstruct the original image sequence (at least coarsely). However, it is also shown that the current reconstruction results based on VM3a are not good enough for many applications.

By combining the enhanced hypothesis generation and consistency check described in Section II.A to the basic approach for volumetric 3-D reconstruction, we obtained another 3-D voxel model for the same image sequence, named “VM5b”. The VM5b contains 86,064 voxels. The average PSNR of the reconstructed 14 images is 19.48 dB with the gain of around 3dB over the average PSNR of the reconstructed images based on the VM3a. We show in Figure 3 the reconstructed images.

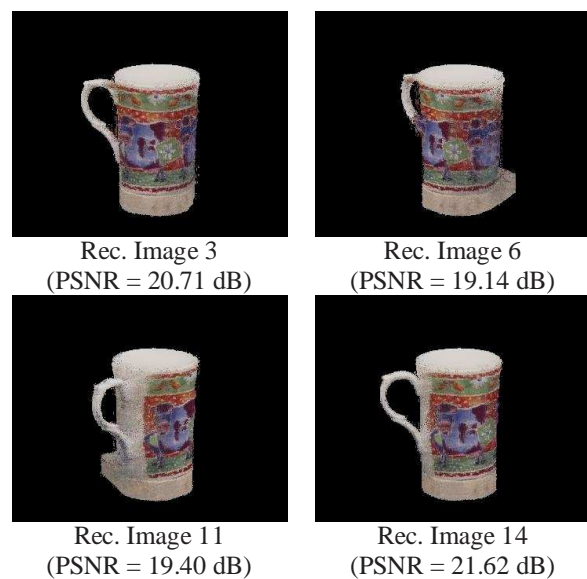


Figure 3 The reconstructed images resulting from re-projecting the VM5b back to image planes for the same camera viewing positions as in Figure 1, as well as the value of PSNR for each reconstructed image.

It is clear that the quality of the reconstructed images is much higher than that of the reconstructed images shown in Figure 2. Many “noisy” pixels outside the true surface of the cup in Figure 2 were removed and the surface of the cup looks smoother. In addition, the size of the VM5b is only around 59 percent of that of the VM3a. This fact indicates that many voxels that were inside the cup have been eliminated from the 3-D voxel model.

By combining the two modifications described in Section II.A and II.B to the basic approach, we obtained the third 3-D voxel model, named “VM5c”. We show in Table 1 the comparison of the data size and the average PSNR of reconstructed images among the obtained three 3-D voxel models.

Table 1 Comparison of data size and the average PSNR 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

Table 1 shows that the VM5c performs the best among the three 3-D voxel models according to both the data size and the quality of reconstructed images. We also show in Figure 4 the reconstructed images corresponding to the images in Figure 1.

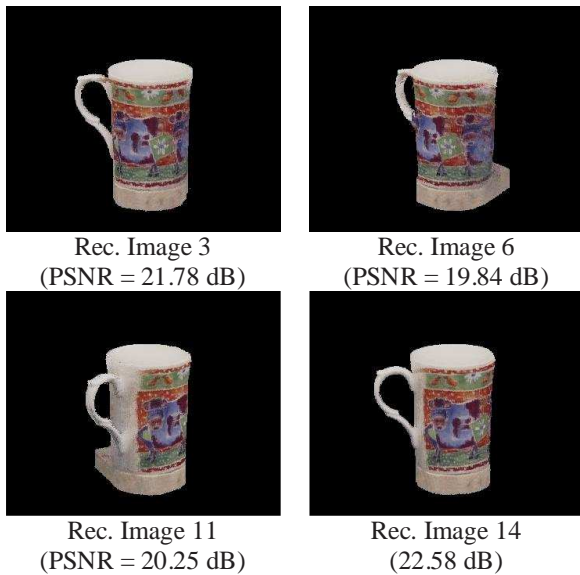


Figure 4 The reconstructed images resulting from re-projecting the VM5c back to image planes for the same camera viewing positions as in Figure 1, as well as the value of PSNR for each reconstructed image.

Synthetic images can even be generated from the obtained 3-D voxel model for new camera viewing positions that are different from those for the original images. Figure 5 shows some synthetic images generated from VM5c.



Figure 5 The generated synthetic images resulting from re-projecting the VM5c back to image planes for new camera viewing positions that are different from those for the original images.

IV. CONCLUSIONS AND DISCUSSION

In this paper, we proposed an algorithm for establishing a 3-D scene voxel model to represent in 3-D space all the available multiple images. Our approach is similar to Eisert’s approach [1][2]. However, we made two modifications to improve its performance: first, we proposed an enhanced hypothesis generation and consistency check; second, we used a new measurement for pixel color information difference based on the physiological characteristics of the human visual system.

We developed three 3-D voxel models for the same image sequence. The first 3-D voxel model (VM3a) is based on the basic approach, which is similar to Eisert’s approach. The second 3-D voxel model (VM5b) is based on the first modification, while the third 3-D voxel model (VM5c) is based on both of the modifications. Experimental results show that the performance of the two 3-D voxel models based on the modified approach (VM5b and VM5c) is significantly better than that of the 3-D voxel model based on the basic approach (VM3a), in terms of the data size of the model and the quality of reconstructed images.

Furthermore, the 3-D voxel model can be combined in multi-view image coding schemes, instead of the commonly used mesh model as well as the texture data. 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.

REFERENCES

- [1] 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.
- [2] P. Eisert, E. Steinbach and B. Girod, "Automatic reconstruction of stationary 3-D objects from multiple uncalibrated camera views", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 261-277, 2000.
- [3] P. Beardsley, P. Torr and A. Zisserman, "3D model acquisition from extended image sequences", *Proc. of European Conference on Computer Vision'1996*, pp. 683-695, 1996.
- [4] M. Pollefeys, R. Loch, M. Vergauwen, and L. Van Gool, "Flexible acquisition of 3D structure from motion", *Tenth IMDSP Workshop'1998*, pp. 195-198, 1998.
- [5] Z. Zhang, Q.-T. Luong and O. D. Faugeras, "Motion of an uncalibrated stereo rig: self-calibration and metric reconstruction", Research Report 2079, INRIA, June 1994.
- [6] <http://www.nt.e-technik.uni-erlangen.de/~eisert/reconst.html>.