Virtual-view-based Stereo Video Composition

Chun-Kai Chang and Hsueh-Ming Hang
Department of Electronics Engineering, National Chiao Tung University, Taiwan
Email: jj2006ace.eecs96@g2.nctu.edu.tw and hmhang@mail.nctu.edu.tw

***Abstract*— Given two sets of videos captured by two sets of multiple cameras, we like to combine them to create a new stereo scene with the foreground objects from one set of video and the background from the other set. We address the camera parameter mismatch and camera orientation mismatch problems in this paper. We propose a floor model to adjust the camera orientation. Once we pick up the landing point (of foreground) in the background scene, we need to adjust the background camera parameters (position etc.) to match the foreground object. The depth information is needed in the above calculation. Thus, new background scenes may have to be synthesized based on the calculated virtual camera parameters and the given background pictures. Plausible results are demonstrated using the proposed algorithms.**

I. INTRODUCTION

The newly developed 3D visual effect offers a whole-new visual experience to human beings, such as 3DTV [1][2], free-viewpoint TV [3] (FTV), and 3D movies. In this study, we like to extend the conventional 2D blue screen (chroma key) to 3D. This work is inspired by the conventional video composition video on Youtube [4], which demonstrates some post-production special effect in the movies and dramas. It shows several video composition techniques such as chroma key and scene completion. The video composition technology allows one to substitute whatever scene he/she wants into another sequence's background as in the video clip [4]. We extend the idea to the stereoscopics by integrating two or more MVD (Multi-view Video plus Depth) sequences into one 3D video. We wish to develop techniques that can produce a new 3D scene in a semi-automatic way. The goal is to make the creation and manipulation of composition as simple and effortless as possible.

The scenario of 3D video composition has some similarities to Augmented Reality (AR). In an AR application, we often need to calculate the camera orientation and movement so that a synthetic object (typically produced by computer graphics) can be properly inserted into a scene. In contrast, we are interested in combining two or more natural videos, where both scenes consist of time-varying natural objects. Because both scenes are not generated by computer, constructing a 3D model of natural objects based on limited views is often difficult (if not impossible). Also, 3D modeling usually requires high computational complexity and a large amount of memory. Here, we adopt the virtual view synthesis technique to generate the background scenes (and sometime foreground objects) that match the scene geometry of the user selected viewpoint.

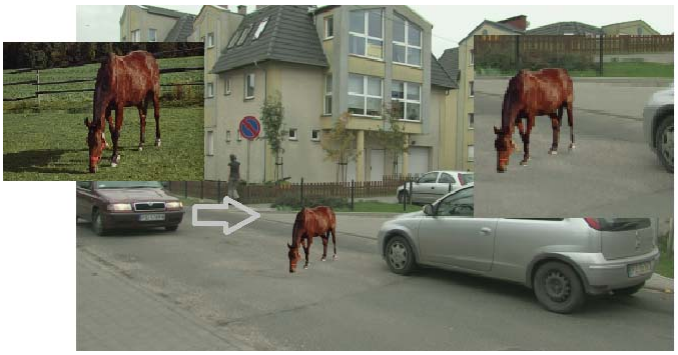

Fig. 1 Conventional composition challenges

The conventional composition such as inserting images into another image with photo realism imposes a number of challenges. The formidable challenges of composition are to satisfy camera geometric consistency and photo content consistency. **Fig. 1** shows a simple composition where we paste a segmented horse (extracted from a scene) into another street scene without any adjustment. There are several noticeable defects in the composite picture. Despite the lack of horse shadow, the first problem you may notice is the size of the horse, which seems to be too small. In addition, the slope of the grass (the horse originally stands on) differs from that of the road. The standing pose of the horse does not seem naturally to match the street floor. Furthermore, if we look at the 3D picture, the depth of house does not match the depth of the ground where it stands.

There are many issues in producing a composed picture. In this study, we will focus on the scene (camera) geometry. In many cases, the true object size and orientation are unknown. Lalonde et al. [5] tackles this problem by proposing an automatically algorithm to estimate camera height and pose with respect to the ground plane in each of the images. Assume that the ground objects stand on is visible and is not tilted from side to side and roughly orthogonal to the image plane, then the camera pose can be estimated based on at least two objects with known heights.

Few researches have combined depth information with composition. Dimitropoulos et al. [6] proposes an approach for 3DTV synthesis. Their work employs the chroma key technique to decompose a scene into foreground and background. They estimate the depth map using two views. Afterwards, they can generate the representation of the scene by combining foreground objects with any background, for both color images and depth maps. However, the direct composition does not consider the mismatch of the

background scenes when captured. And the creativity of new scene is limited by direct background replacement.

The difficulty of 3D video composition mainly lies on the mismatch of the two different scenes. The mismatch type can be mainly divided into "Camera Mismatch" and "Scene mismatch". Camera mismatch refers to the mismatch of the capturing devices. It can be further divided into "Camera Parameter Mismatch" and "Orientation Mismatch". Examples of former are the camera focal length, resolution, principal point offset, signal-to-noise ratio, etc. The latter results from the differences in the camera coordinates, baselines, and orientations, including the initial camera positions and the following movements (zooming, rotation and translation).

The scene mismatch is mainly responsible for perceptive reality. One example is the color temperature mismatch. For instance, the background scene is captured in a sunny day, but the foreground object is pictured in a cloudy day. Another example is the light source, which leads to different shadow or reflection of the object. Our main objective in this study is the adjustment of camera mismatch in two 3D scenes.

II. 3D VIDEO COMPOSITION

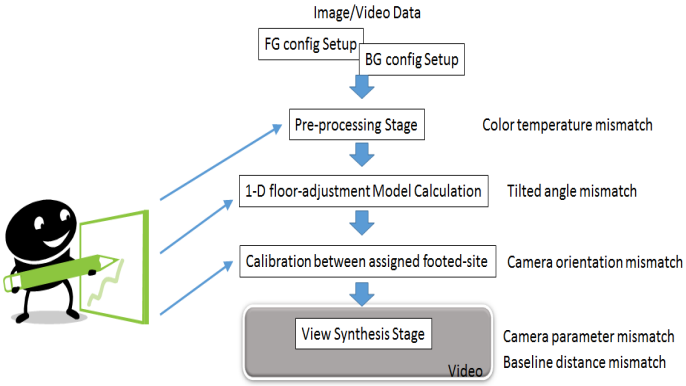

Fig. 2 Overall procedure.

A. System Overview

Fig. 2 shows the flowchart of our proposed 3D composition procedure. The right-side notes annotate the mismatch types to be tackled by each stage (block). At the moment, we assume a person (the creator) is needed to assign key parameters in this procedure such as marking the floor and picking up the landing points (in the background scene) of the foreground objects.

In the rest of this paper, the Target Scene (Foreground Scene) contains the target objects, which are the main objects (Foreground Object) in the composed image. The "Target View" (Foreground View) is the camera view that captures the Target Scene. In most cases, the Target View is adopted as the viewpoint of the Composed Scene. The Background Scene is to be used to replace the background of the Target Scene. We assume the left view is the reference. In other words, the composition starts from the left views (target and background scenes), and the right views are adjusted to match the left views. At the end, a new pair of images (left and right) of the Composed Scene are produced.

B. Preprocessing Stage

The Pre-processing stage includes segmentation and color grading. Limited by space, our focus in this report is geometric adjustment, which is mainly in the floor-adjustment model and the calibration blocks.

C. Camera Orientation Adjustment

Fig. 3 User marks lines on the floor (green line on the left and red line on the right) for calculating the 1-D floor model. The left and right images are Lovebird1 [7] and Poznan_Street [8], respectively (MPEG test sequences).

Fig. 3 shows two scenes captured by two sets of cameras. Their camera orientations are not identical. If we use two persons in the left picture (Lovbird1) as the Target Objects, how do we adjust the orientation of the right picture camera to match that of the left camera? We assume the ground (floor) is our horizontal reference. **Fig. 4** shows the geometric relationship between the floor and the camera orientation. Let the optical axis of the camera be horizontal, and the ground plane has an angle α relative to the optical axis. H represents the distance from optical center to the floor, perpendicular to the optical axis. For each pixel on the marked 1-D line (the green and red lines in **Fig. 3**), z refers to the depth, and θ stands for the angle of a pixel on the image plane relative to the optical axis, and X is the row index along the vertical X-axis on the image plane, whose origin is at the same row of the principal point on the image plane.

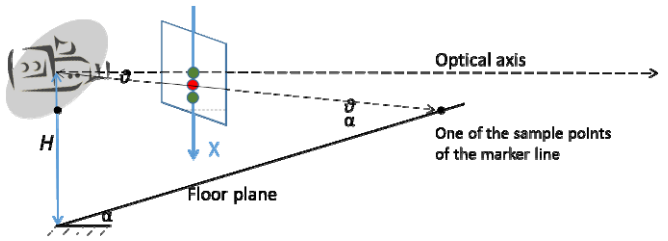

Fig. 4 1D camera orientation vs. floor model.

In **Fig. 4**, we can then observe that

$$z = \frac{H}{\tan\theta + \tan\alpha} = \frac{H}{(\frac{X}{f}) + \tan\alpha} \quad (1)$$

It can be written in the linear form,

$$\frac{1}{z} = (\frac{1}{fH})X + (\frac{\tan\alpha}{H}) = [\omega_1 \quad \omega_0]\begin{bmatrix} X \\ 1 \end{bmatrix} \tag{2}$$

Based the marked 1-D lines on the floor, we can estimate ω by regression in both the foreground and the background scenes, respectively. Each sample of the 1-D line acts as the support vectors for the following regression. Finally, the angle between the floor and the optical axis can be recovered by

$$\alpha = \tan^{-1}(\frac{\omega_0}{\omega_1 f}) \tag{3}$$

Based on the 1-D floor model descried in the above model, we can compensate the camera orientation mismatch in two scenes. Essentially, $R_{fg,bg}(r_x, r_y, r_z)$ is the 3-D rotation matrix between the two camera orientations, where $r_y = r_z = 0$, $r_x = \alpha_{fg} - \alpha_{bg}$. $R_{fg,bg}$ rotates around the X-axis to compensate the mismatch of the orientation based on the common ground floor plane) assumption in two scenes.

Furthermore, the pure rotation changes the depth map because the depth value is the distance projected to the axis. **Fig. 5** shows that the depth value (z_0) of a particular 3D point[1] P changes after a rotation transform $R_{fg,bg}(a,0,0)$ is applied, which gives

$$z' = z_0 \cdot f(\theta, a) \tag{4}$$

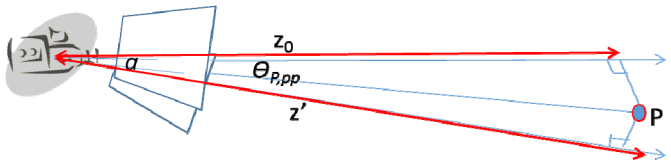

Fig. 5 Geometry relation of pure rotation

$$f(\overrightarrow{\theta_{P,pp}}, a) = \frac{\cos(\left\| \overrightarrow{\theta_{P,pp}} - \begin{bmatrix} 0 \\ a \end{bmatrix} \right\|)}{\cos(\left\| \overrightarrow{\theta_{P,pp}} \right\|)} \tag{5}$$

where z' represents the updated depth value, and z_0 is the original depth value. $\overrightarrow{\theta_{P,pp}} = [\theta_y \quad \theta_x]^T$ is a two-dimensional vector consisting of the angles of the principal point in the y and x directions. and '*pp*' means from the principal point. Note that the depth value of each sequence should be scaled to the same unit. For the MPEG test sequences, the unit information in each sequence can be calculated from the given baseline distance.

D. Camera Orientation Alignment with Assigned Landing Point

The user can place the object landing point at a properly selected location in the background scene. Calibration of the warping parameter for the assigned landing point is as follows.

$$P_0 = \left((fg.K)^{-1} \cdot p_{toe}\right) z_{toe} \tag{6}$$

[1] In this paper, uppercase *P* stands for a 3D point, and the lowercase *p* for a 2D point, in the homogeneous coordinates.

$$P_1 = R_{fg,bg}^T \left((bg.K)^{-1} \cdot p_{landing}\right) z_{landing} \tag{7}$$

$$dt = [dt_x \quad dt_y \quad dt_z]^T = P_1 - P_0 \tag{8}$$

Note that p_{landing} is the 2-D target landing point on the image plane in the background scene, which is often manually picked. It has an associated depth value ($z_{landing}$). And p_{toe} with depth value (z_{toe}) is the "toe" of the Target Object, which is often one of the lowest pixels of the object. $R_{fg,bg}$ is the rotation derived from the previous floor-model. The calibration of the warping parameter aligns the "toe" of the target object with the assigned landing point in the 3D homogeneous coordinate. **Fig. 6** shows several landing points. Once we pick up a landing point, we need to adjust the background scene, so that the depth $z_{landing}$ matches z_{toe} and the position of the two cameras t_{fg} matches t_{bg}. We thus need to move the camera (of the background scene) forward. In other words, the alignment process first estimates the relative position of the foreground camera. The synthesized camera center is determined by the virtual parallel plane above the adjusted background floor located at the obtained relative position. Due to the camera viewpoint shift, the new background scene is synthesized using VSRS. We can see that certain selection of landing points may degrade the image quality of the background due to the view synthesis process.

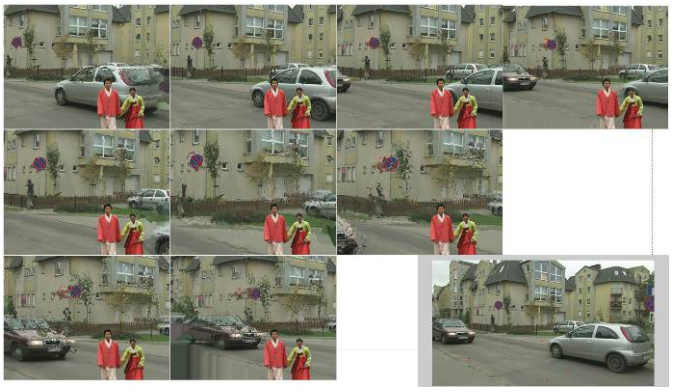

Fig. 6 Some examples of the picked landing points. The camera location needs to be adjusted to match the assigned ground points. The bottom right figure is the background scene. The red dots in the background are the picked ground points.

Additionally, if the assigned landing point is occluded, dr_{toe} is estimated by the 1-D floor model derived from the previous stage to infer a reasonable depth value.

III. VIEW SYNTHESIS

The left view is used to select the landing point and the mark of the 1-D floor plane; then the left background view is synthesized using the rotation and translation matrices derived from the previous stages. The right view of the background scene is regarded as a 1-D parallel view with a baseline distance identical to the foreground stereo pairs. We use VSRS (View Synthesis Reference Software) version 3.5 as

the view generation tool [1]. From the given set of multiple views, we pick up the nearest reference left and right views. An example is shown in **Fig. 7**.

Fig. 7 An example of final stereo composition result. The background is Pozna CarPark [8] (MPEG test sequences).

IV. Experimental Results

We test our camera orientation adjustment in **Fig. 8**. The synthesized pictures with and without the floor-modeling and adjustment are shown. And the 1-D floor-model does improve the photo realism.

Fig. 8 The upper images are the stereo composition results without floor-adjustment; The lower images are with floor-adjustment stage, which seems more natural.

Fig. 9 demonstrates a negative result. In this case, the relative size of the people seems not to conform with the background scene. This is due to the given incorrect depth map as shown in **Fig. 10**, The depth of the marked 1-D floor (red line) in the vertical direction is nearly the same, which leads to the failure of floor-adjustment stage and the wrong depth value of the assigned landing point.

Fig. 9 Composition results using Poznan_Hall2 sequence as the background.

Fig. 10 Texture and depth map of Poznan_Hall2 [8]

V. Conclusions

The goal of the proposed VSRS-based stereo video composition system is to produce good visual quality composite 3-D contents based on the available and calculated geometry information. The challenge is to reduce various types of mismatch between two sets of original 3D scenes. In this study, our focus is on the geometry of the camera and the scenes. We construct a 1-D floor model to aid the camera orientation adjustment. We derive geometric transforms to create new virtual viewpoint so that the new background scenes can be synthesized to match the foreground objects.

VI. Acknowledgment

This work was supported in part by the NSC, Taiwan under Grant 102-2218-E-009-003, NSC 102-2221-E-009-123 and by the Aim for the Top University Project of National Chiao Tung University, Taiwan.

References

[1] P. Kauff, et al., “Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability,” *Signal Process: Image Communication, Special Issue on 3DTV*, pp. 217-234, Feb. 2007.

[2] A. Kubota, et al.,“Multiview Imaging and 3DTV,” *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 10–21, Nov. 2007.

[3] M. Tanimoto, “Free Viewpoint Television (FTV),” *Digital Holography and Three-Dimensional Imaging (DH)*, Vancouver, Canada, June 18, 2007.

[4] http://www.youtube.com/watch?v=a21_WMiTAVE

[5] Jean-François Lalonde, et al., Photo Clip Art, *ACM Transactions on Graphics (SIGGRAPH 2007)*, August 2007, Vol 26. No. 3.

[6] Dimitropoulos, et al., "Improved 3D video synthesis combining graph cuts and chroma key technology," *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2010* , vol., no., pp.1,4, 7-9 June 2010.

[7] G. M. Um, G. Bang, N. Hur, J. Kim, and Y. S. Ho, “3d video test material of outdoor scene,” ISO/IEC JTC1/SC29/WG11, Archamps, France, Tech. Rep. M15371, Apr. 2008.

[8] M. Domanski, et al., “Poznan Multiview Video Test Sequences and Camera Parameters,” ISO/IEC JTC1/SC29/WG11 M17050, Xian, China, Oct. 2009.