3D SCENE RECONSTRUCTION BY MULTIPLE STRUCTURED-LIGHT BASED COMMODITY DEPTH CAMERAS

Jianfeng Wang[*] *Cha Zhang*[+] *Wenwu Zhu*[+] *Zhengyou Zhang*[+] *Zixiang Xiong*[**] *Philip A. Chou*[+]

[*]University of Science and Technology of China
[+]Microsoft Research
[**]Texas A&M University

ABSTRACT

Commodity depth cameras have attracted a lot of research interest recently, in particular the structured-light based Kinect cameras available on the mass market. One important application of such cameras is 3D scene reconstruction and view synthesis. However, a single depth camera often has limited field of view and there is missing depth information when synthesizing a virtual view from a new viewpoint. In this paper, we study the problem of 3D scene reconstruction from multiple structured-light based depth cameras. Since multiple cameras may cause severe interference in the regions where the projected light overlaps, we present a novel plane-sweeping based algorithm to handle such interference. The proposed algorithm takes into account the correlation between multiple projectors and the infrared images as well as the correlation between the infrared images, thereby recovering the depth information for both overlapped and non-overlapped regions. Simulation results demonstrate that the proposed solution is very effective on various scenes.

Index Terms—3D scene reconstruction, multiple depth cameras, structured light.

1. INTRODUCTION

Recently, there have been an increasing number of depth cameras available at commodity prices, such as Microsoft Kinect Sensors [10]. These cameras are active sensors. They emit light (usually in the infrared spectrum) to the environment, and derive the scene's depth information based on structured-light triangulation or time-of-flight measurements. These cameras have created a lot of interesting new research applications, such as 3D shape scanning [5], foreground and background segmentation [4], facial expression tracking [2], etc.

In this paper, we consider the problem of 3D scene reconstruction, which has been an active research topic for decades. It has found many applications including augmented reality, free viewpoint television, and natural user interaction. Traditionally, 3D scene reconstruction was performed with laser scanners or multiple color cameras. The former approach is expensive and slow, and the latter approach is inaccurate, particularly on surfaces where there is no texture. The depth camera provides an alternate, cheap and accurate depth measurement scheme for 3D reconstruction.

Given a depth camera, since the depth information is derived from a single viewpoint, it may contain missing data when viewed from a different viewpoint. A natural solution to the above problem is through the use of multiple depth cameras from different viewpoints, such as the work in [7]. Unfortunately, unlike color cameras that observe the scene passively, active depth sensors emit their own light onto the scene; thus multiple sensors can interfere with each other. Fig. 1 shows an example scene captured by two Kinect cameras. Note when both cameras are turned on simultaneously, the depth quality degrades significantly (Fig. 1 (c)). One must address the interference issue for a setup with multiple depth cameras.

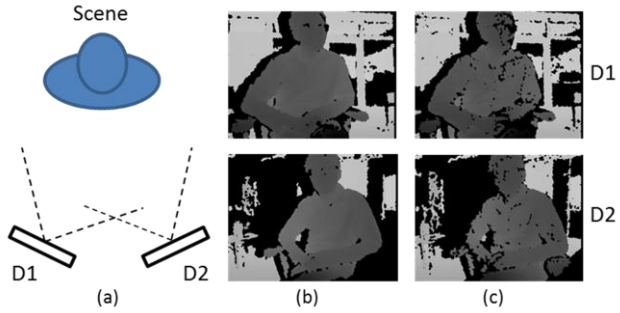

Figure 1. Capturing a scene with two structured-light based depth sensors. (a) System setup. (b) Captured depth images when only one camera operates. (c) Captured depth images when the two depth cameras operate simultaneously. Note the depth images have many more holes.

In [7], to operate multiple depth cameras in the same environment, the authors used three time-of-flight cameras, each operating at a different light modulation frequency. While this is a technically simple solution, customization is required for the depth cameras, which is inconvenient. Another possibility is through time-division multiplexing. That is, different depth cameras operate at different time instances, and thus do not interfere with each other. This solution requires highly accurate synchronization among the cameras, which again is nontrivial to implement. In this paper, we present a novel approach for 3D scene reconstruction from multiple structured-light based depth cameras (SLDC). Unlike existing approaches, we use off-the-shelf cameras directly without modification. A novel depth reconstruction algorithm based on plane-sweeping [3] is proposed, which takes into account the correlation between multiple projectors and infrared sensors as well as the correlation between the infrared images. The approach can therefore recover the depth information for regions with and without interference.

The rest of the paper is organized as follows. The principle of SLDC is briefly reviewed in Section 2. Section 3 describes the proposed algorithm for multiple such cameras, followed by the

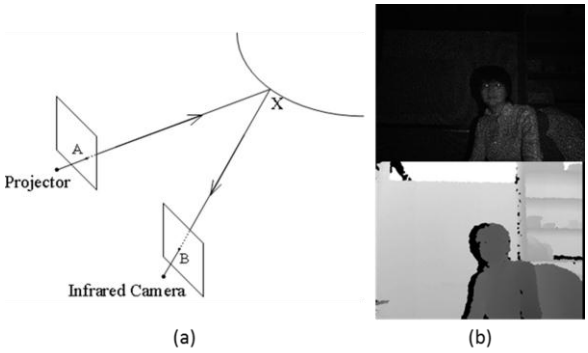

Figure 2. Illustration of a structured-light based depth camera. An infrared projector projects a random pattern onto the scene, which is observed by an infrared camera. Based on triangulation, the depth image can be derived. (a) Setup. (b) Top: image captured by the camera; bottom: depth image.

simulation results in Section 4. In Section 5, we conclude the paper and present the future work.

2. STRUCTURED-LIGHT BASED DEPTH CAMERA

Structured light based stereo vision has been studied in the literature for many years [1]. A typical structured-light based depth camera is constructed by the combination of an infrared projector and an infrared camera. The projector projects a random infrared pattern into the scene, and the camera captures the scene back as images, as shown in Fig. 2. Since the projected pattern is known, the camera can derive the scene's depth information based on standard stereo algorithms such as cross-correlation [6]. The approach is very effective for real-time depth reconstruction, and has been adopted in the popular Microsoft Kinect sensor for creating a new gaming experience.

Since SLDC derives the scene depth assuming the known projected pattern, it is sensitive to external light that may alter the infrared illumination of the scene. In particular, when multiple SLDCs are placed in the same environment with overlapped illumination areas, the depth information derived from each camera may be inaccurate (Fig. 1). Consequently, we face two major challenges when using multiple SLDCs for 3D reconstruction: first, how to correctly distinguish whether a surface area is illuminated by one or multiple projectors; and second, how to perform depth reconstruction in the areas where the projected patterns overlap.

In the next section, we present a plane-sweeping based algorithm that automatically handles the above two challenges.

3. 3D RECONSTRUCTION USING MULTIPLE SLDCS

Consider the general problem of 3D scene reconstruction with *M* projectors and *N* cameras, where *M* and *N* may not be equal (e.g., when additional infrared cameras are added to the setup). We name the projectors $P_1, P_2, \cdots, P_M$ and the cameras $C_1, C_2, \cdots, C_N$. The projectors emit random but time-invariant patterns into the scene, and these patterns are assumed to be known. In addition, we assume that all the cameras and projectors are calibrated beforehand, i.e., for any point $\boldsymbol{X} = (x, y, z)$ in the 3D space, we know how to find its projection onto the 2D images of the projectors and the cameras

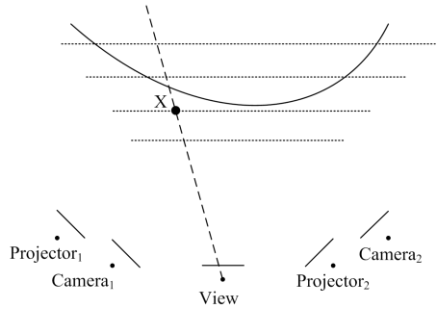

Figure 3. Illustration of the plane sweeping scheme using two structured light based depth cameras.

(the 2D images of the projectors are the patterns themselves):

$$\boldsymbol{x}_{P_m} = \boldsymbol{P}_{P_m}\boldsymbol{X};\ \ \boldsymbol{x}_{C_n} = \boldsymbol{P}_{C_n}\boldsymbol{X}, \tag{1}$$

where $\boldsymbol{x}_{P_m}$ and $\boldsymbol{x}_{C_n}$ are the projected points in the m^{th} projector and the n^{th} camera, respectively. $\boldsymbol{P}_{P_m}$ and $\boldsymbol{P}_{C_n}$ are the projection matrices of the m^{th} projector and the n^{th} camera, respectively.

Given a point $\boldsymbol{X}$ on the surface of the scene, the intensity of the projected 2D pixels shall satisfy two major constraints:

1. Camera Observation Constraint: The cameras shall see the same intensity at the projected pixel, as long as the scene surface point is not occluded:

$$\boldsymbol{I}(\boldsymbol{x}_{C_n}) = \boldsymbol{I}_0. \tag{2}$$

This is the standard multi-view intensity constraint.

2. Projector-Camera Constraint: The observed pixels are the *linear*[1] combination of the projected patterns by the projectors that can illuminate that particular surface point:

$$\boldsymbol{I}_0 = \textstyle\sum_m \alpha_m \boldsymbol{I}(\boldsymbol{x}_{P_m}), \tag{3}$$

where $\alpha_m \geq 0$ are the equivalent reflection ratio for the corresponding pattern.

We develop a simple algorithm based on plane sweeping to perform 3D scene reconstruction using the above two constraints. Plane sweeping is a hypothesis testing scheme illustrated in Fig. 3. Given a virtual view point, the space is sampled into multiple fronto-parallel planes. For a particular light ray, we compute the intersection between the light ray and the test planes, and project the intersection points to the projectors and cameras. The two constraints (Eq. (2) and (3)) are then tested to verify the hypothesis that the surface point is indeed at the intersection. For this purpose, we evaluate the likelihood of the event as below.

3.1. Likelihood of the Camera Observation Constraint

When only the camera observation constraint is considered, the 3D reconstruction problem is a very typical multi-view stereo (MVS) problem. There have been many approaches that can address this problem, ranging from simple plane sweeping based methods [9] to more sophisticated algorithms such as belief propagation [12] and graph cut [8]. In fact, under the multiple structured-light depth cameras setup, the problem is even better conditioned, since the scene surface is textured with the random patterns, thus removing one of the biggest headaches in MVS – textureless surfaces. On the

[1] The linear assumption is usually valid when the number of projectors is small. However, it can happen that the combined illumination may saturate the camera sensors. Considering the nonlinear effects will be our future work.

other hand, when the number of depth cameras is small (e.g., 2-3 depth cameras), MVS may still be insufficient to recover the full 3D depth due to self and mutual occlusions.

In our implementation, we compute the mean-removed cross correlation (MRCC) between corresponding patches to model the likelihood due to the camera observation constraint. Given the hypothesis point X, we project it to all the cameras using Eq. (1). For each projected point $\boldsymbol{x}_{C_n}$, a small surrounding image patch is extracted, denoted as $\boldsymbol{I}_{C_n}$. The MRCC is calculated as:

$$MRCC\left(\boldsymbol{I}_{C_i}, \boldsymbol{I}_{C_j}\right) = \frac{(\boldsymbol{I}_{C_i} - \bar{\boldsymbol{I}}_{C_i})(\boldsymbol{I}_{C_j} - \bar{\boldsymbol{I}}_{C_j})}{\left\|\boldsymbol{I}_{C_i} - \bar{\boldsymbol{I}}_{C_i}\right\| \left\|\boldsymbol{I}_{C_j} - \bar{\boldsymbol{I}}_{C_j}\right\|}, \quad (4)$$

where $\bar{\boldsymbol{I}}_{C_i}$ and $\bar{\boldsymbol{I}}_{C_j}$ are the mean intensity of the patches on camera C_i and C_j, respectively. We take the highest MRCC between the camera pairs as the likelihood of the Camera Observation Constraint for the whole system:

$$L_1 = \max_{i \neq j} MRCC\left(\boldsymbol{I}_{C_i}, \boldsymbol{I}_{C_j}\right). \quad (5)$$

We may also compute the mean patch $\boldsymbol{I}_0$ from the two patches that has the highest MRCC:

$$\boldsymbol{I}_0 = \frac{1}{2}(\boldsymbol{I}_{C_i} + \boldsymbol{I}_{C_j}), \quad (6)$$

which we will use to compute the likelihood of the projector-camera constraint.

3.2. Likelihood of the Projector-Camera Constraint

The projector-camera constraint takes more sophistication to explore, as the linear weights α_m in Eq. (3) are unknown. In fact, α_m at least depends on the distance from the surface point to each projector, the surface orientation, and self/mutual occlusions. Given the mean patch $\boldsymbol{I}_0$ obtained in the previous subsection, we solve a least square fitting problem as:

$$\{\hat{\alpha}_m\} = \arg\min_{\{\alpha_m\}} \left\|\sum_m \alpha_m \boldsymbol{I}_{P_m} - \boldsymbol{I}_0\right\|^2, \quad (7)$$

where $\boldsymbol{I}_{P_m}$ is a patch surrounding the hypothesized intersection's projection to projector $\boldsymbol{P}_m$. Such a fitting problem can be easily solved using the pseudo inverse.

The likelihood of the projector-camera constraint is thus computed as the MRCC of the fitting result with $\boldsymbol{I}_0$:

$$L_2 = MRCC\left(\sum_m \hat{\alpha}_m \boldsymbol{I}_{P_m}, \boldsymbol{I}_0\right). \quad (8)$$

The overall likelihood of a hypothesized intersection is thus computed as:

$$L = \frac{1}{2}(L_1 + L_2). \quad (9)$$

3.3. Practical Implementation

In real-world systems, the number of structured-light depth cameras adopted is usually small. Consequently, the baseline between the depth cameras can be large. It is well known that multi-view stereo with wide baselines is a very challenging problem due to occlusions and perspective projections. On the other hand, although multiple depth cameras can interfere with each other, due to the random patterns adopted in such cameras, each camera still has some capability of determining the correct depth under interference (as demonstrated in Fig. 1 (c)).

This leads to a simpler but more efficient scheme as follows. We first let each depth camera obtain depth values independently. This is usually performed by the depth cameras' hardware, thus does not incur any computational cost on the computer. The depth images may contain holes in the interference regions, since the hardware algorithm has a hard threshold to ensure all reported depth values are correct. Given a virtual viewpoint, we warp the depth images to the desired viewpoint, and then fill the holes through maximizing the likelihood as in Eq. (9). This algorithm works very well, as shown in the next section.

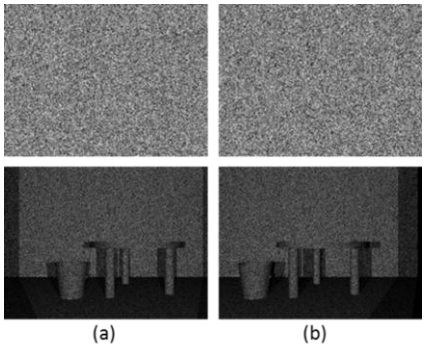

Figure 4. Illustration of the synthetic scene experiments. Top: random patterns of the simulated projectors. Bottom: "captured" images. The brighter region in the bottom images are the regions with the two patterns overlapped.

4. EXPERIMENTAL RESULTS

To validate the performance of the proposed method, we conduct experiments using three synthetic scenes, *teapot*, *vase* and *table/bucket*, rendered by the popular ray tracing software POV-Ray[2] [11]. For each scene, two depth cameras are used to simultaneously capture the depth of the scene. Each depth camera contains a projector and a camera. The projector is simulated with a point light source centered in a cube. Five faces of the cube are solid, and the other face is modulated with a pseudo-random pattern. The camera is placed 7.2 cm away from the projector, and captures the scene. The two depth cameras are about 30 cm apart. The random pattern and two example images captured by the infrared cameras are shown in Fig. 4.

Fig. 5 (a) and (b) shows the depth images reconstructed using cross-correlation based stereo method for each depth camera independently. Due to interference from the other camera, both depth maps have holes since the maximum MRCC score between the captured image and the known projector pattern for those light rays is below a fixed threshold 0.5. In Fig. 5 (c), we attempt to reconstruct the depth map at the center viewpoint between the two depth cameras using the higher MRCC of the two cameras. It can be seen that some of the holes are filled, though the depth map still has relatively poor quality. In Fig. 5 (d), we reconstruct the depth map at the center viewpoint using MVS by maximizing Eq. (5) for each light ray during plane sweeping. The result is better in the overlapped regions but worse around occlusion boundaries. Fig. 5 (e) shows the result of the proposed method (Section 3.3). It can be seen that the resultant depth map is very good, similar to the ground truth depth (Fig. 5 (f)).

We further measure the peak signal to noise ratio (PSNR) of the reconstructed depth in Table 1. It can be seen that the proposed method performs much better than direct depth merge or MVS.

[2] We tried to use the Kinect sensor, and found that the depth values reported by the sensor are very poorly calibrated across different units. Future work is necessary to calibrate them accurately before testing the proposed method.

Table 1. PSNR of reconstructed depth maps in dB. Note the results of MVS are much worse than the other two because MVS cannot reconstruct depth for the areas not seen by both cameras.

	Depth merge Fig. 5 (c)	MVS Fig. 5 (d)	Proposed Fig. 5 (e)
Teapot	22.0746	12.9473	33.6717
Vase	26.6831	13.6862	32.1119
Table/bucket	20.3470	12.8033	27.5648

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a novel 3D depth reconstruction algorithm using multiple structured-light based depth cameras. The algorithm fuses the results from structured-light based stereo and multi-view stereo, and is capable of combating the interference caused by multiple active sensors. The solution can be applied to commodity depth cameras without special customization, and is thus attractive in practice. Future work includes the calibration of the returned depth values from commodity depth sensors, and applying the proposed method in real world examples.

6. REFERENCES

[1] Batlle, J., Mouaddib, E. and Salvi, J., "Recent progress in coded structured light as a technique to solve the correspondence problem: a survey," *Pattern Recognition*, Vol. 31, No. 7, pp. 963—982, 1998.

[2] Cai, Q., Gallup, D., Zhang, C., and Zhang, Z., "3d deformable face tracking with a commodity depth camera," in *ECCV*, 2010.

[3] Collins, R. T., "A space-sweep approach to true multi-image matching," *CVPR* 1996.

[4] Crabb, R., Tracey, C., Puranik, A., and Davis, J., "Real-time fore-ground segmentation via range and color imaging," in *CVPR Workshop on ToF-Camera based Computer Vision*, 2008.

[5] Cui, Y., Schuon, S., Chan, D., Thrun, S., and Theobalt, C., "3D shape scanning with a time-of-flight camera," in *CVPR*, 2010.

[6] Faugeras, O. *et. al*, "Real-time correlation-based stereo : algorithm, implementations and applications," INRIA technical report #2013, 1993.

[7] Kang, Y.-S. and Ho, Y.-S., "High-quality multi-view depth generation using multiple color and depth cameras," *ICME* 2010.

[8] Kolmogorov, V. and Zabih, R., "Multi-camera Scene Reconstruction via Graph Cuts," *ECCV 2002*.

[9] Kutulakos, K. and Seitz, S., "A theory of shape by space carving," *IJCV*, Vol. 38, No. 3, pp. 199—218, 2000.

[10] Microsoft, http://www.xbox.com/en-us/kinect/.

[11] POV-Ray, http://www.povray.org/

[12] Sun, J., Zheng, N.-N. and Shum, H.-Y., "Stereo Matching Using Belief Propagation," *IEEE Trans. On PAMI*, Vol. 25, No. 7, pp. 787-800.

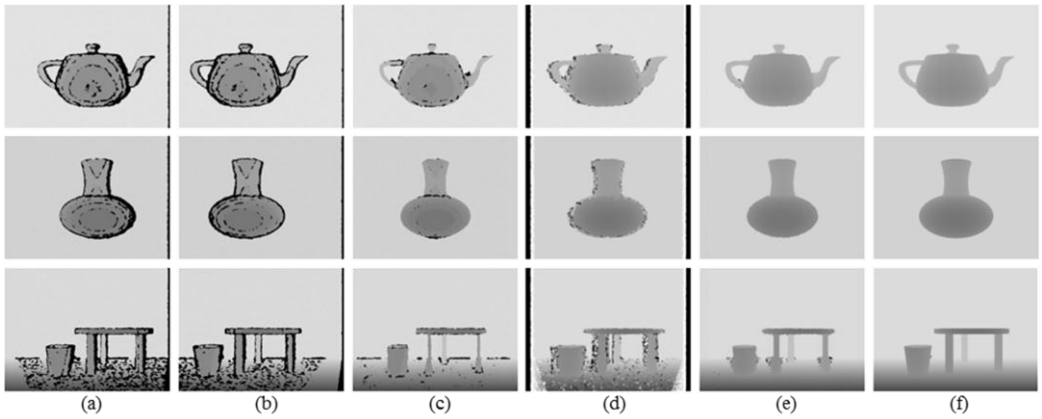

Figure 5. Experimental results on 3 synthetic scenes simulated with POV-Ray. From top to bottom: teapot, vase, table/bucket. (a) Depth image "captured" from the left camera. (b) Depth image "captured" from the right camera. (c) Merged depth map rendered at a center viewpoint. (d) Depth reconstructed at a center viewpoint using multi-view stereo based on maximizing Eq. (5). (e) Depth reconstructed at a center viewpoint using the proposed method. (f) Ground true depth map at the center viewpoint.