INDOOR MAPPING BASED ON RGB-D AND DSLR CAMERAS

Tzy-Shyuan Wu[1] and Fuan Tsai[2]
[1]Department of Civil Engineering, National Central University, Jhongli City, Taoyuan 32001, Taiwan,
Email: shyuan_wu@hotmail.com
[2]Centre for Space and Remote Sensing Research, National Central University, Jhongli City, Taoyuan 32001, Taiwan,
Email: ftsai@csrsr.ncu.edu.tw

KEY WORDS: photogrammetry, Kinect, SFM reconstruction

ABSTRACT: RGB-D cameras, which capture both RGB images and per-pixel depth information, recently became a popular indoor mapping tool in the field of computer vision. One of the mainstream solutions for indoor mapping and modeling is to create 3D point cloud from multiple images. However, the major drawback of image-based approaches is the lack of points extracted in featureless areas. The integration of RGB-D based sensors and cameras may fill up these voids in featureless areas and create a uniformly distributed point cloud of indoor environments. In this research, a hardware consisting of Kinects and digital single-lens reflex (DSLR) cameras is assembled and the data processing procedure for integrating these two kinds of devices to generate 3D point cloud is established. There is interference between Kinects; hence the field of view of the Kinects cannot overlap with one another. Thus, DSLRs are used to bridge the Kinects and provide a more accurate ray intersection condition, which takes advantage of the higher resolution and image quality of the DSLR cameras. Bundle adjustment is used to resolve the exterior orientation (EO) of all RGB images acquired by Kinect and DSLR. The EO of Kinect at each frame is used as an initial value to combine these point clouds at each frame into the same coordinate system. The result shows that the design of the hardware and the data processing procedure can generate dense and fully colored point clouds of indoor environments even in featureless areas.

1. INTRODUCTION

Indoor environment modeling has been extensively developed in recent years. Three-dimensional (3D) indoor mapping aims to create a digital representation of the environment. Once the rich and high-precision digital model is constructed, the information of an environment is preserved. In addition, it can be utilized in reconstruction, measure, indoor location positioning and other applications.

These days, RGB-D sensors became a popular indoor mapping tool in the field of computer vision because of the low-cost and its capability to capture color images along with per-pixel depth information in the same time. One kind of RGB-D camera, Microsoft Kinects, was used in this study. In computer vision, the most used technique in 3D reconstruction of Kinect is Kinect Fusion. However, Kinect Fusion is based on camera tracking (Newcombe et al., 2011). It depends on depth variation in the scene. Scenes must have sufficient depth variation in view to be able to track successfully (Microsoft, 2014). That is, if there is a smooth wall or less depth changes places, the Fusion may fail to construct the model. To overcome this shortcoming, this study developed a photogrammetry and structure-from-motion approach to realize the indoor mapping using RGB-D and DSLR cameras.

In photogrammetry, one of the mainstream solutions for indoor mapping and modeling is to create 3D point cloud from multiple images. In this case, multiple images connect each other by means of features. After feature extraction, 3D point clouds can be created by intersecting feature points on multiple images. However, this approach is less effective in the featureless places.

Both RGB-D camera and images have their advantages but limitations. To make up a deficiency, this research combines RGB-D camera's information and the high-resolution images captured by DSLR camera. The integration of these two kinds of devices may fill up these voids in featureless areas and create a uniformly distributed point cloud of indoor environments. SFM reconstruction and bundle adjustment are used to obtain the exterior orientation (EO) of all RGB images acquired by Kinect and DSLR. The EO of Kinect at each frame is used as an initial value to combine these point clouds at each frame into the same coordinate system.

2. METHODOLOGY

2.1 Procedure

Figure1 shows the overall procedure of the developed method. In this study, RGB-D camera (Microsoft Kinect) and digital single-lens reflex (DSLR) camera is used to capture the information of the indoor environment. Kinect provides RGB images and per-pixel depth values. According to the information it captured, point cloud of each frame can be generated. On the other hand, a DSLR camera captures multiple high-resolution images of the indoor environment. A visual structure from motion system (VisualSFM) is used to link the relationship of all the images captured by both Kinect and DSLR camera. In this part, a sparse point cloud can be created in a photogrammetry way. And bundle adjustment is used to resolve the exterior orientation (EO) of all images. Those exterior orientations are used as the initial values to combine these point clouds at each frame into the same coordinate system. Finally, colored point clouds of indoor environments are generated. The quality of the combined point cloud can be evaluated by comparing the width, height and length of identified objects in the model against in situ measurements. Another verification choice is to compare the ground control points' coordinates which were set in the indoor environment.

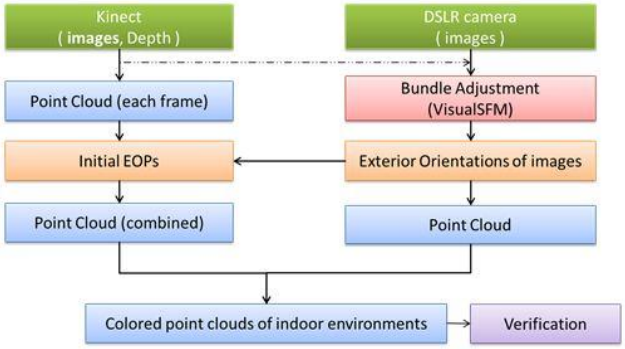

Figure 1: General procedure of the developed method

2.2 RGB-D Camera: Kinect

RGB-D cameras capture RGB images along with depth information. Different kinds of RGB-D cameras use different ways to obtain the depth information. Two typical types of traditional RGB-D cameras are time-of-flight (TOF) cameras and stereo cameras. A TOF camera resolves distance based on the known speed of light, measuring the time-of-flight of a light signal between the camera and the subject for each point of the image. A stereo camera obtains the depth value by the method of photogrammetry. It captures two images in one time, making parallax measurements on the object to the camera. A new and developing RGB-D camera is Kinect, which was initially used as an input device by Microsoft for Xbox game console. Microsoft Kinect can provide color and depth images synchronously. Kinect uses the technique of Light-coding. The sensor launches a laser speckle and captures the coding light back from the scene to calculate depth values. Comparing those three kinds of RGB-D cameras, Kinect has much lower cost than traditional ones and are more widely used in recent years (Han et al., 2013).

Microsoft Kinect has a RGB camera and 3D depth sensors. 3D depth sensors consist of an infrared CMOS camera and infrared projector. It captures 640*480 or 1280*960 pixels in color images and 640*480 pixels in depth images at up to 30 frames per second (Han et al., 2013). Figure 2 shows an example frame captured by Kinect. Based on the RGB image and depth information, colored point cloud (Figure 3) can be generated for every frame. However, because of the different captured range of RGB and IR camera, it is necessary to obtain the correct corresponding pixels before point cloud generation.

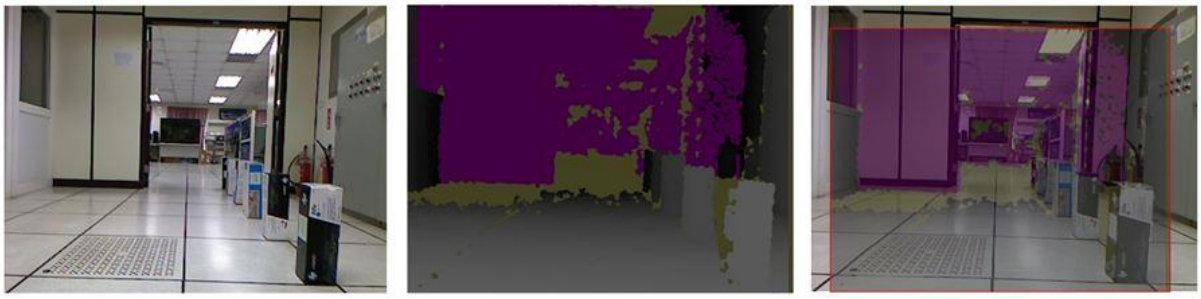

Figure 2: RGB (left) image and depth (middle) image captured by Kinect at a resolution of 640*480 pixels and the corresponding ranges of RGB and Depth image (right)

Using Kinect for 3D mapping, there are some drawbacks from its limitation. One is that it can only provide limited

distance in depth information. The available distance is about 1 to 4 meters. The spatial and depth resolutions are millimeter to centimeter depending on the range. The spatial (x, y) resolution is about 2 to 20 millimeters, and distance (z) resolution is about 1 to 70 millimeters. In addition, the random error of depth measurements increases with increasing distance from the sensor, and reaches about 4 centimeter at the maximum range (Khoshelham and Elberink, 2012). Figure 3 shows an example of one frame point cloud. In this example, the limitation of the capture depth value was 4095mm.

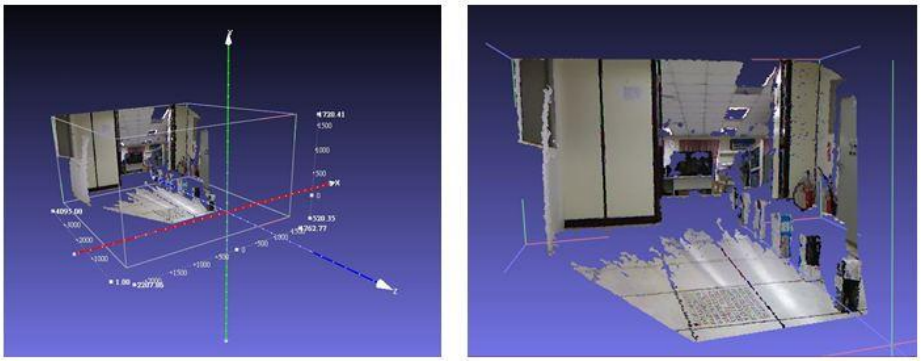

Figure 3: (Left) an example of point cloud (unit: millimeter; 307200 points); (Right) a close look of the point cloud

In Figure 3, it also can be found that Kinect cannot determine the depth value of glass objects, such as fluorescent light tubes and windows. Moreover, if the tilt angle is too large, the depth value may not be stable. After testing more than one place with overlaps, it is found that the depth image also has interference, and this may cause the captured depth value instable. Therefore, to proceed with Kinect-based indoor mapping, it is necessary to consider Kinect's properties and also its limitations.

2.3 SFM Reconstruction

For image-based 3D reconstruction, color images taken from DSLR camera and Kinects are processed with a Visual Structure from Motion System (VisualSFM), to compute relative points of images. The processing steps include: a) feature extraction (default SIFT processing); b) feature matching; c) sparse reconstruction; d) bundle adjustment; e) dense matching defaults. First of all, lens distortions can be solved in this processing. After bundle adjustment, exterior orientation of each images captured from Kinects and DSLR camera can be resolved (Wu, 2011).

DSLR camera captures high-resolution photographs, which provide detail information of the environment. However, if the scenes are featureless or the information is not adequate, it is difficult to perform feature extraction and feature matching. This may result in lacks of point clouds in the featureless places. To overcome this disadvantage, Kinects' 3D point cloud can be used to complete the model.

2.4 Exterior orientation

In the part of SFM reconstruction, at least four ground control points are used to define the model coordinate in an object space coordinate system. After the SFM reconstruction, the Kinect images' camera positions and exterior orientations are computed. The exterior orientation of Kinect at each frame is used as an initial value to combine these point clouds at each frame into the same coordinate system. Table 1 is the information obtained after the SFM reconstruction.

Table 1: The information obtained after the SFM reconstruction

No.	Camera parameter information
1	Focal Length (of the undistorted image)
2	2-vec Principal Point (image center)
3	3-vec Translation T (as in P = K[R T])
4	3-vec Camera Position C (as in P = K[R -RC])
5	3-vec Axis Angle format of R
6	4-vec Quaternion format of R
7	3x3 Matrix format of R
8	[Normalized radial distortion] = [radial distortion] * [focal length]^2
9	3-vec Lat/Lng/Alt from EXIF

3. EXPERIMENT

An experiment was performed using the proposed procedure. Figure 4 shows the indoor gallery chosen as the study area in this research. Several stickers were set on the walls as ground control points that can be used to verify the results.

Figure 4: Study Area

3.1 Data Source

Data acquisition included two parts, Kinect and DSLR camera. Table 2 shows the information of the DSRL camera and Kinect used in this experiment. Two Kinect sensors (Microsoft Kinect for windows) were used to capture two sides of the walls, with 12 frames per second at the same time. For the high resolution photographs, this test acquired 250 images in total in this environment. Photographs were captured including straight and transverse side. In front of the wall, there is about 80% overlap between adjacent images. Figure 5 displays an example of the overlap situation. The images used in the SFM reconstruction included 250 high resolution photographs captured by DSLR camera and 24 RGB images captured by Kinect.

Table 2: Information of the sensors

Sensors	Information	
Kinect	Version	Microsoft Kinect for windows
	Capture range	Standard range motion capture (0.8 – 4m)
	Depth image resolution	640*480 (pixels)
	Color image resolution	1280*960 (pixels)
	Frequency	12 frames per second
DSLR camera	Focal length (fc)	[2730.7509 2729.7600] ± [1.0439 0.9437] (pixel)
	Principal point (cc)	[2845.5841 1846.1958] ± [1.1031 0.9738] (pixel)
	Distortion (kc)	[0.0082 -0.0063 0.0028 0.0024 0.0000] (pixel)
	Pixel error	[0.5045 0.2869] (pixel)

Figure 5: there are about 80% overlaps between captured images

3.2 Result and Discussions

3.2.1 Results of SFM Reconstruction

In this part, VisualSFM (V0.5.25) is utilized for the 3D reconstruction of all the images. Figure 6 shows the 3D points after the sparse reconstruction. This figure also concludes the camera position of the images. The smaller images in

the middle are the RGB images captured by Kinects. It can be seen that there are more points in the feature-rich places, like the poster on the wall. Figure 7 is the result of 3D dense points after the dense reconstruction. There are fewer points in the featureless places on the flat walls and the ceiling because those places do not have enough features in multiple images for feature matching.

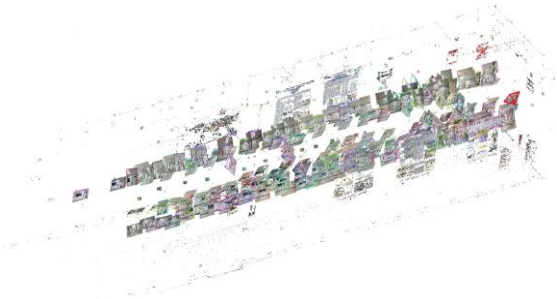

Figure 6: Sparse reconstruction result

Figure 7: Dense reconstruction result

3.2.2 Results of indoor mapping point cloud

Point clouds of the Kinect per frame's coordinates are adjusted according to the exterior orientation parameters solved from Visual SFM. Figure 8 are point clouds combined from four frames of Kinect data. Point clouds are in the same coordinate after EO transformation. Then, combine Kinect's point cloud with the SFM reconstruction result in the same coordinate (Figure 9). As displayed in this figure, there are still some orientation problems and differences. In this case, adjust the Kinect point cloud manually to the ideal places (Figure 10). Figure 11 shows a close look of the combined point cloud after adjustment. In order to display the result more clearly, just one side of the Kinect point cloud is shown in the figure.

Figure 8: (Left) Left-side wall's Kinect point cloud and (Right) Right-side wall's Kinect point cloud after EO transformation

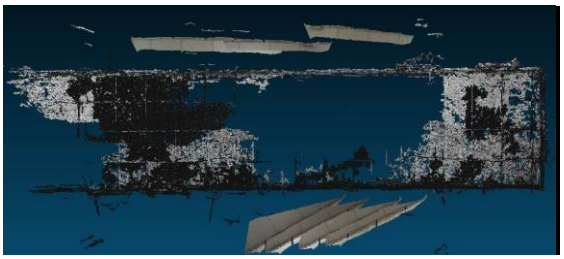

Figure 9: Combined point cloud after EO transformation

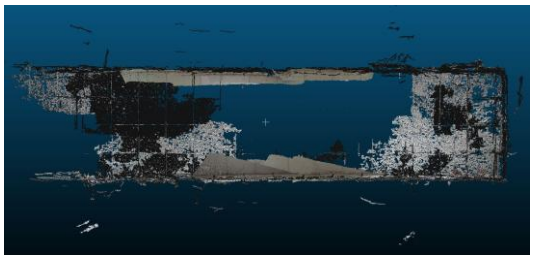

Figure 10: Combined point cloud after adjusted manually

Figure 11: one side of the ideal combined dense point cloud model (unit: meter)

3.3 Problems and Discussions

Because this experiment only captured depth images using two Kinects at the same time, the obtained data were not enough for reconstructing a complete model. According to the presented results, the coordinate transformation of the Kinect may still have difference from the point cloud generated from SFM reconstruction. This problem may be addressed by taking Kinects' local coordinate system into account and improving the transformation of EO parameters. Moreover, iterative closest point (ICP) algorithm may be used to refine the initial EO and resolve the scale factor to generate a more accurate 3D point cloud (Henry et al., 2012; Yue et al., 2014). Nevertheless, the example demonstrates that the developed method can effectively combine the data acquired using these two devices to create a dense and fully colored point cloud of an indoor environment.

4. CONCLUSIONS

The proposed two kinds of sensors have their own properties and drawbacks. Although only capturing small range and lower resolution RGB images, Kinect can provide a real distance and scale. On the other hand, DSLR camera can capture high-resolution images. However, if the environment is featureless, the SFM reconstruction may fail. This research integrates computer vision and photogrammetry approaches to combine these two indoor mapping devices for indoor mapping. The proposed method uses the exterior orientation parameters to transfer the Kinect sensors' point clouds coordinates into the same coordinate system as the dense reconstruction results of DSLR images. Using the proposed method, a dense and fully colored point cloud of indoor environments can be generated effectively and accurately even in featureless areas using RGB-D and DSLR cameras.

REFERENCES

Han, J., Shao, L., Xu, D. and Shotton, J., 2013. Enhanced Computer Vision with Microsoft Kinect Sensor: A Review. IEEE Transactions on Cybernetics, pp.1318-1334.

Henry, P., Krainin, M., Herbst, E., Ren, X. and Fox, D., 2012. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. The International Journal of Robotics Research, 31(5), pp. 647-663.

Khoshelham, K. and Elberink, S., 2012. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications, Sensors, 12(2), pp.1437-1454.

Microsoft. Retrieved August 20, 2014, from http://msdn.microsoft.com/en-us/library/dn188670.aspx

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S. and Fitzgibbon, A., 2011. KinectFusion: Real-time dense surface mapping and tracking. Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on, pp.127-136.

Wu, C., 2011. VisualSFM: A Visual Structure from Motion System, Retrieved August 20, 2014, from http://ccwu.me/vsfm/

Yue, H., Chen, W., Wu, X. and Liu, J., 2014. Fast 3D modeling in complex environments using a single Kinect sensor, Optics and Lasers in Engineering, 53, pp.104-111.