RECONSTRUCTION OF INDOOR MODELS USING POINT CLOUDS GENERATED FROM SINGLE-LENS REFLEX CAMERAS AND DEPTH IMAGES

Fuan Tsai[a*], Tzy-Shyuan Wu[b], I-Chieh Lee[a], Huan Chang[b] and Addison Y. S. Su[c]

[a]Center for Space and Remote Sensing Research
[b]Department of Civil Engineering
[c]Research Center for Advanced Science and Technology
National Central University, Zhong-li, Taoyuan 320 Taiwan
ftsai@csrsr.ncu.edu.tw

KEY WORDS: Indoor modeling, RGB-D, Kinect, SFM reconstruction, Point clouds

ABSTRACT:

This paper presents a data acquisition system consisting of multiple RGB-D sensors and digital single-lens reflex (DSLR) cameras. A systematic data processing procedure for integrating these two kinds of devices to generate three-dimensional point clouds of indoor environments is also developed and described. In the developed system, DSLR cameras 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. Structure from Motion (SFM) reconstruction is used to link and merge multiple Kinect point clouds and dense point clouds (from DSLR color images) to generate initial integrated point clouds. Then, 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 using Helmert (seven-parameter) transformation. Experimental results demonstrate that the design of the data acquisition system and the data processing procedure can generate dense and fully colored point clouds of indoor environments successfully even in featureless areas. The accuracy of the generated point clouds were evaluated by comparing the widths and heights of identified objects as well as coordinates of pre-set independent check points against in situ measurements. Based on the generated point clouds, complete and accurate three-dimensional models of indoor environments can be constructed effectively.

1. INTRODUCTION

Three-dimensional modeling of indoor environments is an emerging topic in the researches and applications of building modeling, indoor navigation, location-based services and related fields in recent years. One of the common objectives in three-dimensional indoor mapping is to create a digital representation of the environment. Once the rich and high-precision digital model is constructed, the complete and accurate information of an indoor environment is preserved and can be used for a variety of applications.

A popular and conventional approach of creating three-dimensional indoor models is to generate three-dimensional point clouds from multiple digital images. In this case, multiple images must be connected with each other based on identified features and three-dimensional point clouds can be created by intersecting feature points on multiple images. However, the major disadvantage of this image-based approach is the lack of points extracted in featureless areas and therefore, it may be less effective in areas such as plain walls or long, simple corridors from which few features will be identified and extracted for model reconstruction. As technology advances, new equipments and softwares are developed and have great potentials to achieve better and more effective reconstruction of three-dimensional indoor models. Among the newly developed instruments, RGB-D cameras (such as Microsoft Kinect) have attracted attentions from researchers in the fields of computer vision and photogrammetry. This type of camera can capture both RGB images and per-pixel depth information simultaneously. Although the effective range of data acquisition using Kinect sensors is short (1 to 4 meters), they can be utilized as a new tool for indoor mapping and model reconstruction (Han et al., 2013; Yue et al., 2014). However, the most popular three-dimensional reconstruction method for Kinect data is using Kinect Fusion, which is based based on camera tracking. It depends on depth variation in the scene. Scenes must have sufficient depth variation in view to be able to track successfully, so Kinect Fusion may fail to construct the models in places where depth changes less significantly (Newcombe et al., 2011). Never the less, the integration of RGB-D based sensors and digital cameras may fill up the voids in featureless areas and create uniformly distributed points cloud of indoor environments. Accordingly, this research constructed a mobile data acquisition system consisting of multiple Kinects and digital single-lens reflex (DSLR) cameras and developed a systematic data processing procedure for integrating these two kinds of devices to generate three-dimensional point clouds of indoor environments.

2. MATERIAL AND METHOD

The developed data acquisition system consists of up to four RGB-D (Kinect) cameras and four DSLR cameras mounted on a steel rack, which in term can be installed on a pull-cart or similar platform. Figure 1 displays an example of the developed mobile data acquisition system and sample images from the RGB-D and DSLR cameras. The RGB-D cameras used in this study are Microsoft 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 with other types of RGB-D cameras, Kinect has much lower cost than traditional ones and is more widely used in recent year (Han et al., 2013). Table 1 shows some basic parameters of Kinect.

There are some drawbacks when using Kinect for 3D mapping from its limitation. First of all, it can only be used in indoor envi-

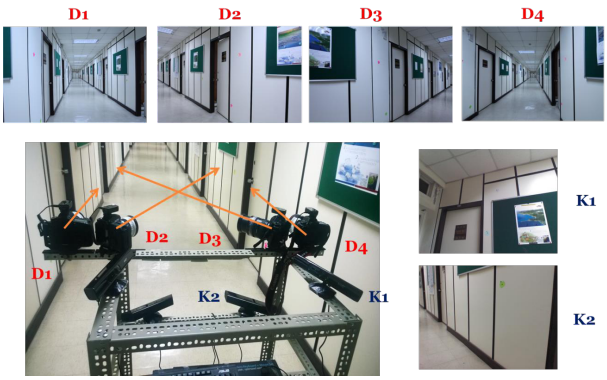


Figure 1: The developed mobile data acquisition system with multiple RGB-D and DSLR cameras.

Table 1: Basic parameters of Kinect

Range	0.8-4 m	
Image Size	Depth	640x480
	Color	1280x960
Frequency	12 fps	
Accuracy	Spatial	2-20 mm
	Distance	1-70 mm

ronment. Secondly, it can only provide limited distance in depth information (08-4 m). Therefore, it may not be appropriate to acquire data in a large, open space environment. 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).

Kinect provides RGB images and per-pixel depth values. According to the information it captures, a 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. The general procedure of VisualSFM is listed in Algorithm 1. This process generates a sparse (merged) point cloud in a photogrammetry way. Also, feature points of each frame were extracted to be used as tie-points. 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, 3D point clouds of Kinect can be used to complete the model. According to the extracted feature points, Kinect point clouds of each frame can be transformed into the same coordinate system with dense point clouds through Helmert (7-parameter) transformation as illustrated in Fig. 2. Finally, colored point clouds of indoor environments are generated. The quality of the combined point cloud can be evaluated by comparing against the coordinates of pre-set ground control points.

3. EXPERIMENTAL RESULTS

The developed mobile RGB-D and DSLR data acquisition system was deployed to reconstruct three-dimensional models of a long corridor in a campus building. The results are compared with pre-set control and check points to evaluate the accuracy

Algorithm 1 General procedure of VisualSFM

1. Feature extraction (default: SIFT)
2. Feature matching
3. Sparse reconstruction
4. Bundle adjustment
5. Dense matching

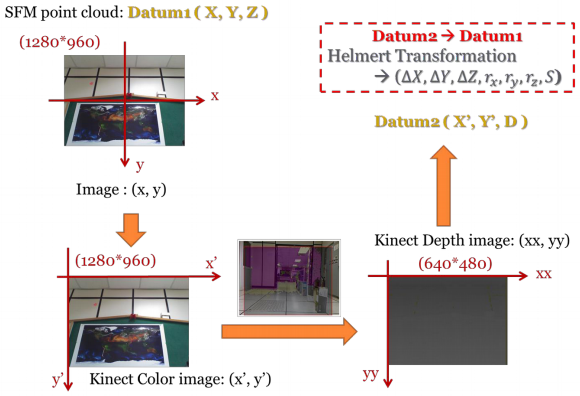


Figure 2: Determining Helmert 7-parameters from feature points

of the generated point clouds. The first experiment utilized two Kinect sensors to capture two sides of the walls at a speed of 12 frames per second. A total of 250 high resolution digital images were also captured by DSLR cameras at the same time. The high resolution photographs covered both straight and transverse sides. In front of the wall, there is about 60% overlap between adjacent images. The images used in the GCP-based SFM reconstruction included 250 high resolution photographs captured by DSLR camera and 24 RGB images captured by Kinect. The second experiment utilized four Kinect RGB-D sensors and four DSLR cameras in the same environment. In this test, 118 photographs per DSLR camera and 100 color images per Kinect were used in the SFM reconstruction.

Figure 3 displays the result of dense point clouds reconstructed from the DSLR photographs and color images of Kinect sensors. As mentioned previously, despite the high resolution of DSLR images, there are few feature points in flat, feature-less regions, thus resulting in many holes in the reconstructed point clouds. These voids can be augmented by merging the point clouds derived from Kinect depth images into the SFM generated point cloud using the Helmert transformation in order to produce a more complete model. The effect is more obvious in the second experiment as shown in Fig. 4, which displays the original (SFM reconstructed) dense matching point cloud and the merged point cloud result (combination of DSLR, Kinect color images and depth images). Figure 5 shows an inside view of the reconstructed hallway point cloud. The example in Fig. 4 and 5 clearly demonstrates that merging the data of depth images into the SFM dense matching result significantly reduces the voids in the reconstructed point cloud model, especially in the feature-less regions.

The reconstructed point cloud models were compared with pre-set ground control points and field-measurement performed with total station equipment for accuracy assessment. Table 2 lists the accuracy assessment result of the the dense matching point cloud in X, Y, Z directions and the distance. From the table, it can be seen that the average errors in the three axises are less than the 1

Figure 3: Point cloud generated from DSLR and Kinect color images

cm where the average distance measurement error is about 0.86 cm. Table 3 further evaluate the accuracies of the point clouds from two Kinect sensors (K1 and K2) as indicated in Fig. 1. Similarly, the average measurement errors in the three axises are also less than 1 cm. However, the standard deviation values are higher than Table 2 and the distance errors are 1.62 cm and 1.61 cm, respectively. This is reasonable because the depth sensor in Kinect is less accurate than the GRB camera, thus resulting in larger uncertainty in the measurement. Nevertheless, the results are still very accurate and the accuracy should be adequate for indoor mapping applications.

Table 2: Accuracy assessment of dense matching point cloud (unit: cm)

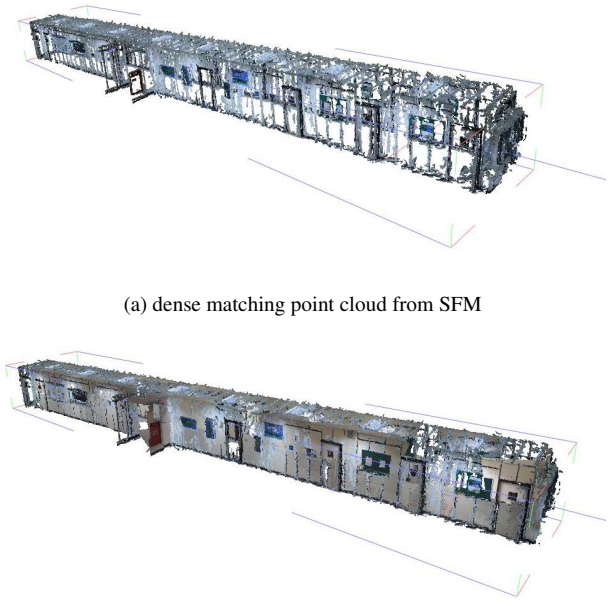
	dX	dY	dZ	Distance
MEAN	-0.05	0.61	0.45	0.86
STDDEV	0.60	0.70	0.21	

Table 3: Accuracy assessment of merged point cloud (unit: cm)

		dX	dY	dZ	Distance
K1	MEAN	0.23	-0.27	-0.20	1.62
	STDDEV	1.01	0.81	1.08	
K2	MEAN	-0.44	0.08	-0.12	1.61
	STDDEV	0.95	1.06	1.04	

4. CONCLUSIONS

This research proposes an approach to integrate the point clouds generated from DSLR and RGB-D (Kinect) cameras. In spite of short range and relatively low resolution color images, Kinect can provide real distance and scale information. The proposed method employs Structure From Motion (SFM) algorithm to integrate the DSLR and Kinect color images to generate a dense matching point cloud. However, the reconstruction may fail in feature-less regions, resulting in voids or holes in the generated point clouds. These voids are augmented by merging the point clouds from depth images into the reconstructed model using 7-parameter transformation. Using the proposed method, dense and fully colored point clouds of indoor environments can be generated effectively and accurately even in featureless areas. Experiment results indicate that the distance error of the constructed merge point clouds is less than 2 centimeters using the proposed processing procedure. This process will be an effective approach to build an indoor mapping model.



(a) dense matching point cloud from SFM

(b) merged point cloud

Figure 4: Comparison of original dense matching and merged point clouds

Figure 5: Inside view of the constructed point cloud model

ACKNOWLEDGEMENT

This study was supported, in part, by the Ministry of Interior and the Ministry of Science and Technology of Taiwan (ROC) under project numbers SYC1040120 and MOST-103-2221-E-008-076-MY2, respectively.

References

Han, J., Shao, L., Xu, D. and Shotton, J., 2013. Enhanced computer vision with microsoft kinect sensor: A review. IEEE Transactions on Cybernetics 43(5), pp. 1318–1334.

Khoshelham, K. and Elberink, S., 2012. Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors 12(2), pp. 1437–1454.

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. In: 10th IEEE Symposium on Mixed and Augmented Reality (ISMAR2011), pp. 127–136.

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.