INDOOR 3D MAPPING BY KINECT SENSOR

Vaibhav Katiyar[1] and Masahiko Nagai*[2]
[1] Graduate student. Remote Sensing and GIS, Asian Institute of Technology,
Paholyothin Highway, Klong Luang, Pathumthani 12120; Thailand; Tel: +66827072968
E-mail: vaibhav.katiyar@ait.ac.th

[2] Associate Director, Geo-informatics Center(GIC), Asian Institute of Technology,
Paholyothin Highway, P.O. Box 4, Klong Luang, Pathumthani 12120, Thailand; Tel: +6625245599
E-mail: nagaim@ait.ac.th

KEY WORDS: 3D Mapping, Indoor Mapping, Kinect Sensor, RGB-D Sensor

ABSTRACT: In geospatial domain, mapping plays an important role in planning and management and when it comes to high accuracy then the best solution is 3D mobile mapping. However different environments make it difficult to generate 3D with proper scale and high level of accuracy, for example making 3D of a room where some people are moving. One another common problem is expensive hardware and complex software that limits 3D reconstruction as common practice in real life.
This study is mainly concerned about constructing indoor 3D by using Kinect sensor. Kinect sensor was launched by Microsoft as a part of XBox-360 but later they released "Microsoft research Kinect SDK" to work with Kinect as a standalone device. Kinect sensor gets the images in color (RGB) and depth. Corresponding points in the successive frames of RGB and Depth images can be detected using SURF and RANSAC. These corresponding points are mostly corners of some object or some clear mark which is clearly visible in both the images. In succession 3D transformations can be applied on space co-ordinates of resultant corresponding points, to get a common co-ordinate system. However in some of the cases if the object is moving, consider the case of a moving person in a room, and then the finding common co-ordinate system is not as easy as above. This research mainly concern outliers due to humans, moving in room that's why it used human detection methods from depth images. 3D points (outliers) corresponding to that object is removed from the actual data to get more accurate 3D. Aim of this study is to make 3D of a common practice in day-to-day life, within limited amount of money expenditure and achieving high end of accuracy.

1. INTRODUCTION

Detailed 3D map of an indoor environment provides rich information about its geography and orientation which make ease in navigation, telepresence and semantic mapping. Also 3D map is very useful in emergencies such as it can provide good information to the fire fighters in the case of fire in a building. Many researches are already in progress in the field of 3D indoor mapping using different techniques like Laser scanners (Liu et al., 2010), and Hybrid Range Finders (Morris & Dryanovski, n d). Some of these techniques are very expensive and some are very complex. Which is difficult to be used by a non-technical person. Now-a-days RGB-D cameras (e.g. Kinect) are available at very low cost, which can be used without requiring any technical knowledge or training. So anyone can afford these sensors and play with them. 3D reconstruction is not only concerned about location but also contextual information. Accurate information of the location can be easily computed from the 3D point cloud with the help of depth camera whereas for contextual information, color camera is used which works on visible light.
In this study, we present a prototype for automated 3D generation by using Kinect sensor which has RGB and depth camera that makes life easy to reconstruct 3D of the interior of a building even in bad light conditions. SURF (Bay et al., 2008) algorithm is applied for finding corresponding points and RANSAC (Fischler and Bolles, 1981) for removing outliers. Although resolution of Kinect camera is not very good (640x480 for color and 320x240 for depth image) but its high frame rate makes it very useful for 3D development in real time. Sometime while making 3D of interior of a building, a common problem is a moving person in front of the camera, which is just a noise. This study uses human detection algorithm and then those points (noise) is removed from the original point cloud data, does making 3D reconstruction easy in different environmental conditions and more robust.

2. SYSTEM CONFIGURATION

2.1 Sensor and SDK

As shown in Figure 1 Kinect sensor is the combination of three types of sensors Visual sensor, Audio sensors, and Motor sensor. Main focus of this study is concerned about Visual sensor which consists of one RGB camera and one IR/Depth camera. RGB video stream resolution was 640x380 and Depth stream resolution was 320x240 with the is

Figure 1. Kinect Sensor

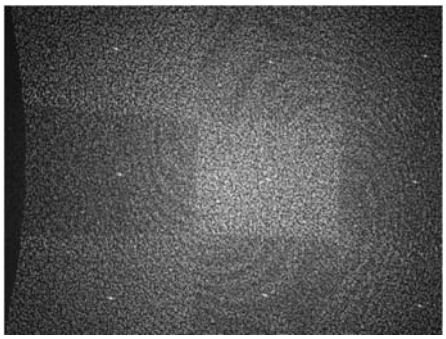


Figure 2. Kinect IR projection sample pattern (Reprinted from www.ros.org)

output rate of 30 frames per second, this frame rate plays an important role in real time 3D mapping with dense point cloud. The angular FOV was 57° horizontally and 43° vertically but by using the tilt motor we can get +-27° in the vertical direction.

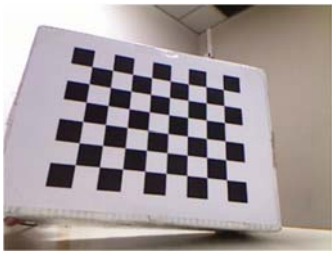
Kinect finds depth with the help of IR projector and IR CMOS camera. Kinect projector projects a known pseudo random pattern (a sample is shown in Figure 2) on objects, at the same time IR camera observes the scene and then by using the principle of triangulation depth of every point can be determined.

Microsoft launched "Microsoft Kinect SDK beta" to work with data streams of Kinect on 16th June 2011. The SDK can handle visual, audio, and motor sensors. Currently this SDK is only compatible with Windows7. This SDK includes drivers, API's, and other documents to provide an environment where developer can develop some application using C#, C++, VB on Visual Studio 2010.

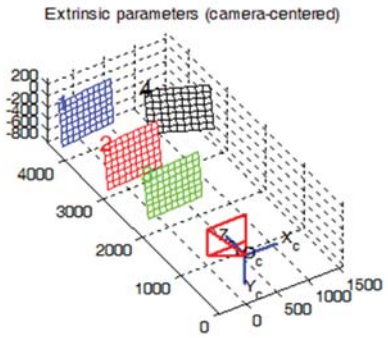
2.2 Calibration and Setting

For calibration, at first Kinect should be kept in a well lightened room on hard surface so its angle does not get change and can capture good RGB images. The object of interest should be at distance in between 0.8 to 4 meter from the camera; beyond this range depth camera will return a fixed value which can be considered as no value. Calibration technique can be classified into two categories: photogrammetric calibration and self-calibration (Zhang, 2000) Due to unavailability of any calibration object, in this research study self-calibration technique is used. We used Zhang camera calibration technique which can be divided into the following steps (Zhang, 2000):

1) Detecting feature point in image (Used chessboard as in Figure 3(a))
2) Estimating intrinsic and extrinsic parameters (Figure 3(b)) using the closed form solution.
3) Estimating the coefficient of radial distortion by solving linear least square.
4) Refining the complete set of parameters by minimizing.



(a)



(b)

Figure 3. (a) Corner detection of chessboard for using it as feature point. (b) Shows orientation of images with respect to fixed camera.

3. DATA PROCESSING

Complete work flow of data processing and generation of 3D primarily consist 6 steps as shown in Figure 4:

Step1. Camera Calibration is one of the beginnings and most important step because quality of data depends upon how accurately cameras are calibrated. Zhang camera calibration is used for this step.
Step2. This step concerned with point cloud. Locations of each point are estimated by depth image which is then integrated with texture information retrieved from color image.
Step3. Finding corresponding points by applying SURF on 2D frames and then applying RANSAC on 3D points. These 3D points is corresponding to SURF feature point retrieved from point cloud data
Step4. Common co-ordinate system is determined using Helmert's 3D transformation and then all points in same co-ordinate system are converted accordingly.
Step5. Noise and errors introduced in previous steps have no visible affect in short distance but for a long run it can creates a big problem, loop closing detection technique is used to remove these errors. In this method camera returns on initial position and then by back propagation cumulative error is removed.
Step6. After getting consistent point cloud we only concerned to visualize this dynamically which is implemented in OpenSceneGraph(OSG).

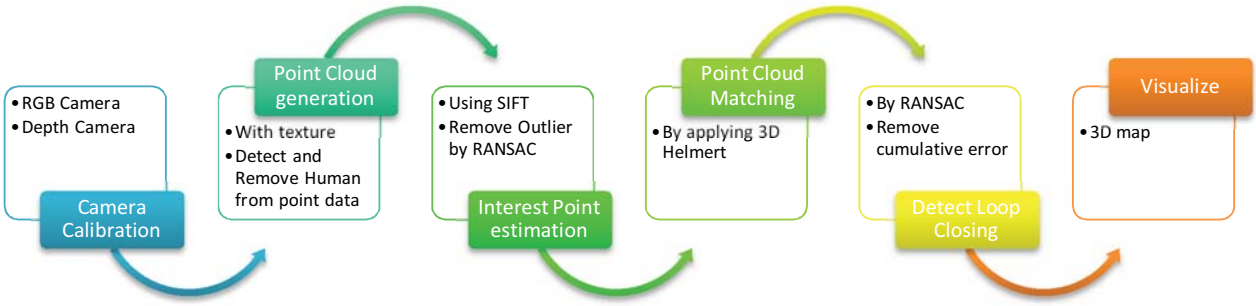
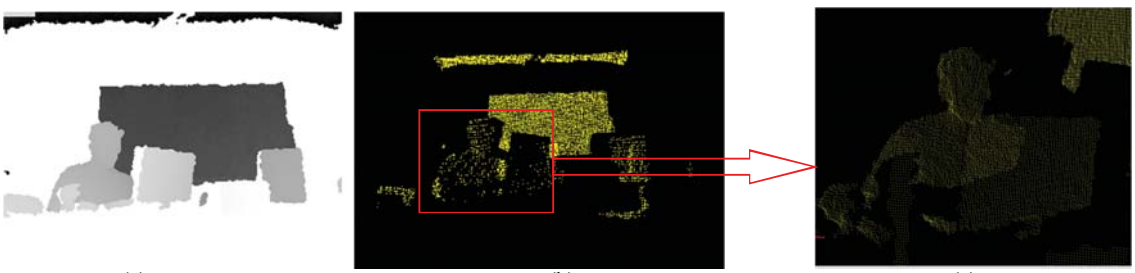


Figure 4. Work-Flow of study

3.2 Point cloud generation

Kinect IR camera returns depth image (Figure 5(a)) which is just the measured distance from its center to the object. We find depth (Z value) for each point and then a text file is generated having X, Y, Z co-ordinates for each frame. Points stored in text file can be visualized as shown in Figure 5(b) & (c). In general Kinect frame per second is very high (~30fps) but we used 1 frame per second to reduce the amount of data collected. When Kinect takes depth image, RGB camera works simultaneously and captures the same scene. It is easy to find texture information from the corresponding RGB image and store with depth information.
In this step human detection algorithm also runs to locate human body and to remove all points which refer to it. This ensures removal of any noise from final point cloud caused by the people who come in front of the camera during 3D generation.



(a) (b) (c)

Figure 5. (a)Depth image (b) & (c)Point cloud

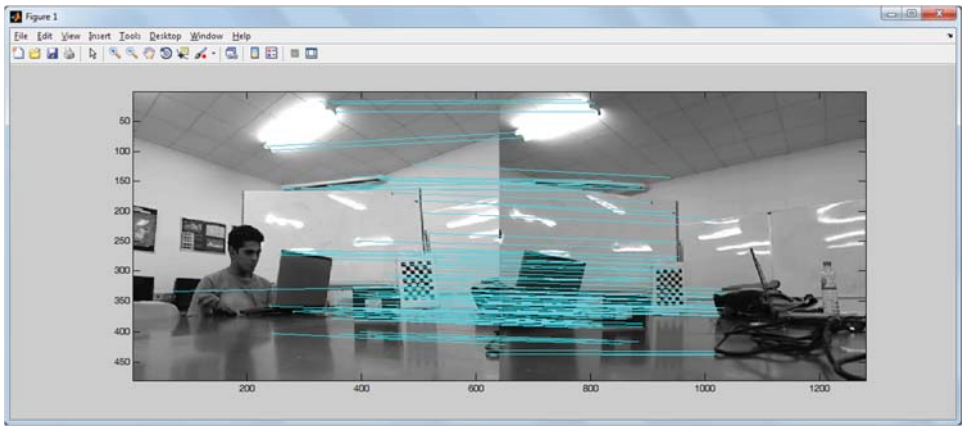
3.3 Interest point estimation

SURF detector and descriptor is a newly introduced concept and it is proved that its performance is either equal or superior in comparison to other existing methods like SIFT, PCA-SIFT moreover computational efficiency is also

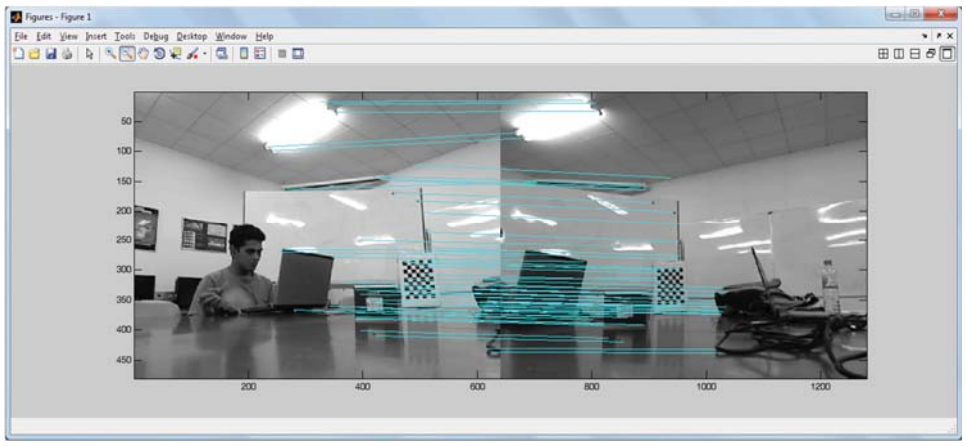
significantly better with aforementioned method (Pimenov, n d). According to Herbert Bay, SURF is less sensitive to noise in comparison to SIFT because SURF integrates the gradient information within a sub-patch unlike SIFT which depends upon orientation of individual gradient (Bay et al., 2008). Real time processing is affected by so many noises hence, in this study SURF was used for feature point estimation.

(a)

(b)



(c)



(d)

Figure 6. (a) & (b) Test images (c) SURF features point in between (a) and (b). (d) Corresponding points after applying RANSAC over SURF feature points

After getting SURF feature points (Figure 6(c)), 3D co-ordinate of corresponding points can be found from the stored point cloud data. RANSAC is applied on the resultant 3D points. RANSAC procedure tries to fit plane on these 3D points and remove outliers. The output of this procedure will provide pure corresponding points on both the images.

3.4 3D transformation for point cloud matching

The 3D Helmert transformation is a very old and reliable technique for producing transformation between two different Cartesian co-ordinate systems. After Appling 3D Helmert transformation (by equation 1) on all SURF points (corresponding points) of two frames which remain after removing outliers, a rotational and translational matrix is obtained as a resultant. Using both of these matrixes all points are transformed into the same co-ordinate system.

$$X_{transform} = T + sRX \quad (1)$$

Where

$X_{transform}$ is transformed co-ordinate vector.
T is translational or shift vector.
s is scale factor (In this study it will be always 1).
R is rotational matrix consist three rotational vector r_x, r_y, r_z on each axis x, y, z respectively.
X is initial co-ordinate vector.

3.6 Loop Closing detection with the help of RANSAC

Transformation is a very good technique for tracking camera position with time in between successive frames. However, there is small error present either in transformation, calibration, quantization of depth values, or some other noise which is not notable in short range but in long run it will be significant (Figure 7(a)). Loop closing detection is one of the important concepts for removing such kind of cumulative errors.
RANSAC is one of the efficient procedures to find loop closing. RANSAC can be run on SURF feature points of all previous frames for detecting the loop closeness, but it is very time consuming process. So for removing this complexity, we used a "keyframe" list. For making this list we used same RANSAC procedure in successive frames. As long as inliners between successive frames are not in between threshold, the frame is added as new keyframe. As camera continues to move and at some point of time when there are not enough inliers, then that frame will be next keyframe. At this stage, RANSAC is applied between this new keyframe and all other previous keyframes. A closer is detected if enough 3D features points, returned by RANSAC, matches (Henry, Krainin, Herbst, Ren, & Fox, n d).
After close loop detection total cumulative error is collected. This error is then back-propagated making the 3D model more accurate (Figure 7 a, b, c).

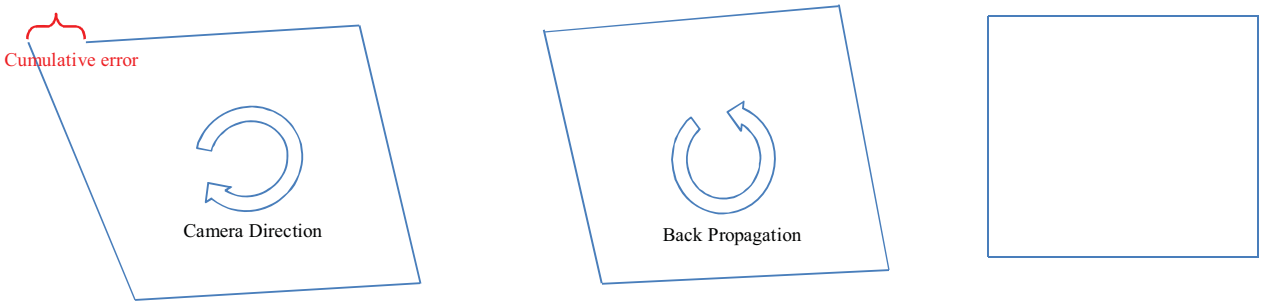


Figure 7. (a) Close loop detection with cumulative error (b) Back propagate and remove error (c) Original path/loop without error

3.7 Point Cloud Visualization

For visualizing point cloud data this research used Open Scene Graph (OSG). OSG is one of the good open sources rendering library. It can handle big amount of data which is very important requirement in this research study because it continuously receives frames containing millions of point. One another important feature of OSG is that real time processing is easy and we can add points dynamically.

4 CONCLUSIONS

Building 3D model of indoor environments has so many applications like navigation, robotics, gaming and others. In the field of 3D indoor reconstruction many technologies already exist like Laser scanning, Vision-only approaches,

and Hybrid range finders but some of them are too expensive, require lot of computation or not robust in different kind of environment like limited lighting, lack of distinctive feature.

In this study we used SURF with RANSAC for finding corresponding points which results good number of points without outliers. Noise and error of every step is easily removed by loop closing technique. One of the achievements of this study is detecting human and removing them from the scene without affecting 3D making procedure, it makes indoor 3D reconstruction more robust.

From this research study, we realize that in the field of indoor 3D-mapping RGB-D cameras like Kinect can really provide reliable results with high accuracy. Currently Kinect has some limitation like its SDK is only compatible with Windows 7 and some hardware constraint like range of an object should be in between 0.8 meters to 4 meters. But this is just a beginning; in the near future we will have more robust, compatible and highly accurate RGB-D sensor which will make 3D mapping a child's play.

5 REFERENCES

5.1 References from Journals:

Bay, H., Ess, a, Tuytelaars, T., & Vangool, L., 2008. Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding, 110(3), 346-359. doi: 10.1016/j.cviu.2007.09.014.

Fischler and Bolles, 1981 M.A. Fischler and R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (6) pp. 381–395.

Liu, T., Carlberg, M., Chen, G., Chen, J., Kua, J., & Zakhor, A. (2010). Indoor Localization and Visualization Using a Human-Operated Backpack System. Image Processing, (September), 15-17.

Zhang, Z., 2000. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11), 1330-1334. doi: 10.1109/34.888718.

5.2 References from Other Literature:

Henry, P., Krainin, M., Herbst, E., Ren, X., & Fox, D. (n.d.). RGB-D Mapping : Using Depth Cameras for Dense 3D Modeling of Indoor Environments. Intel Lab.

L. Xia, C.-C. Chen, and J. K. Aggarwal, Human Detection Using Depth Information by Kinect, International Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR (HAU3D), Colorado Springs, CO, June 2011.

Morris, W., & Dryanovski, I. (n.d.). 3D Indoor Mapping for Micro-UAVs Using Hybrid Range Finders and Multi-Volume Occupancy Grids. Sensors (Peterborough, NH).

Pfister, H., Ý, M. Z., & Ý, M. G. (n.d.). Surfels : Surface Elements as Rendering Primitives. Methods, 335-342.

Pimenov, V. (n.d.). Fast Image Matching with Visual Attention and SURF Descriptors. Control.

5.3 Reference from sites:

http://www.vision.ee.ethz.ch/~surf/index.html

http://research.microsoft.com/en-us/um/redmond/projects/kinectsdk/about.aspx

http://primesense.com

http://channel9.msdn.com

http://www.ros.org