Stereo and Kinect Fusion for Continuous 3D Reconstruction and Visual Odometry

Ozgur Yilmaz[#*1], Fatih Karakus[*2]

[#]*Department of Computer Engineering, Turgut Özal University*
Ankara, TURKEY
[1]ozyilmaz@turgutozal.edu.tr

[*]*Aselsan Inc., MGEO Divison*
Ankara, TURKEY
[2]fkarakus@aselsan.com.tr

***Abstract*— Robust and accurate 3D reconstruction of the scene is essential for many robotic and computer vision applications. We are proposing a system solution that can accurately reconstruct the scene both indoor and outdoor, in real-time. The system utilizes both active and passive visual sensors in conjunction with peripheral hardware for communication, and suggests a significant accuracy improvement over state-of-the-art SLAM algorithms via stereo visual odometry integration.**

***Keywords*— SLAM, Stereo, Kinect, Visual Odometry, ICP.**

I. INTRODUCTION

Many robust solutions have been proposed in the recent years for 3D reconstruction of a scene using active (Kinect, ToF cameras) or passive (stereo) visual sensors [1]-[7]. The application is also known as SLAM (Simultaneous Localization and Mapping) in robot vision literature [8] or SfM (Structure from Motion) in computer vision studies [9], in which visual tracking and registration is utilized to estimate the camera motion and build a 3D map of the environment at the same time. More recent work emphasized the importance of low error in camera motion estimation [6],[10], dense reconstruction [3], and reconstruction of extended areas [11]-[15].

A specific brand of RGB-D camera named Kinect® gave a boost in SLAM studies due to its specifications and low price. Specifically, KinectFusion algorithm [3] was a huge step towards real-time operating 3D reconstruction systems given its reconstruction accuracy and speed. It uses a volumetric representation of the scene called truncated sign distance function (TSDF), and fast iterative closest point (ICP) algorithm for camera motion estimation. KinectFusion is an orthogonal approach to interest point/pose estimation based algorithms [6], [16]. It optimizes 3D model detail and real-time performance but trades-off other dimensions: registration accuracy and 3D model size. Usage of depth image based registration technique (ICP) causes large errors when camera motion is large or scene is poor in 3D structure (i.e. flat regions). And voxel based scene representation is problematic for reconstruction of a large area due to memory limitations. In recent stdies, KinectFusion algorithm is modified and extended to include solutions to these shortcomings.

Fig. 1 A. Kinect plus Bumblebee XB3 stereo rig used in the system. B. Padded shoulder image stabilizer for image acquisition. Brand name RPS [26]

For improving registration accuracy, better energy minimization procedures were defined for ICP [17], [18]. Alternatively, RANSAC based visual matching and motion estimation was used as an initialization [19] for ICP which avoids converging to local minima or it was used for sanity check [15]. Several remedies have been proposed to extend the area of reconstruction [11]-[15]. The proposed approaches are mainly based on automatically detecting if the camera moves out of the defined volume and re-initiating the algorithm, after saving the previously reconstructed volume as TSDF [13],[14] or saving into a more efficient representation [11], [15]. Most recent work [15] proposes a complete solution to both registration and volume extension problems. However, their system is limited with the capabilities of the Kinect

sensor: only indoor operation and only IR projection based depth map.

The problem we tackle is different from the ones studied in recent literature: robust multi-session 3D reconstruction for both indoor and outdoor operation. For some applications, due to limited energy resources and memory capacity, there is no need to reconstruct a very large region as a whole but some disconnected areas of interest in the region; executing reconstruction in multiple sessions. KinectFusion framework is a very good candidate for fine reconstruction of the disconnected areas, but disconnected models need to be located in a global coordinate system for holistic visualization. Also, Kinect sensor is not suitable for operation under sunlight, and needs a stereo depth image support. We propose a stereo plus kinect hybrid system that utilizes visual feature based stereo odometry for navigation in the scene and kinect+stereo depth image for 3D reconstruction. The proposed system: 1. Fuses Kinect and Stereo depth maps in order to be able to work under both low light and sunlight conditions 2. Uses stereo visual odometry navigation solution to stabilize fast ICP used in KinectFusion framework 3. Keeps track of relative transformation between multiple KinectFusion 3D models using stereo visual odometry. Therefore we are using stereo for both improving KinectFusion reconstruction under sunlight and for continuous visual odometry. Nonetheless, visual odometry is utilized for both aiding fast ICP in KinectFusion and for locating multiple 3D models with respect to each other.

The main contributions of our study are: introduction of stereo in KinectFusion framework, utilization of stereo visual odometry for improving registration and providing global localization, design of multi-session 3D reconstruction concept, and a complete system solution to 3D reconstruction.

II. System and Approach

A. Hardware

We are proposing a complete system solution for large area 3D reconstruction, both for indoor and outdoor operation, in any terrain. The system consists of a Kinect+Stereo (Bumblebee XB3®) rack (Fig. 1A) for imaging, a padded shoulder image stabilizer (Fig. 1B) for ergonomics and a laptop (IEEE 1394b express card installed) for Bumblebee image acquisition and wireless image transfer. The system enables mobile acquisition of Kinect (with 12V battery power supply) and stereo images even in rough terrains. The images are uploaded to a workstation through wireless transfer, and processed.

B. Algorithmic Approach

Stereo is an essential aid to Kinect for outdoor operation and we are introducing StereoFusion and Kinect+StereoFusion. Stereo depth image [20] is used instead of Kinect depth image in the former and the two depth maps are fused in the latter. To our knowledge, this is the first time stereo is utilized in KinectFusion framework.

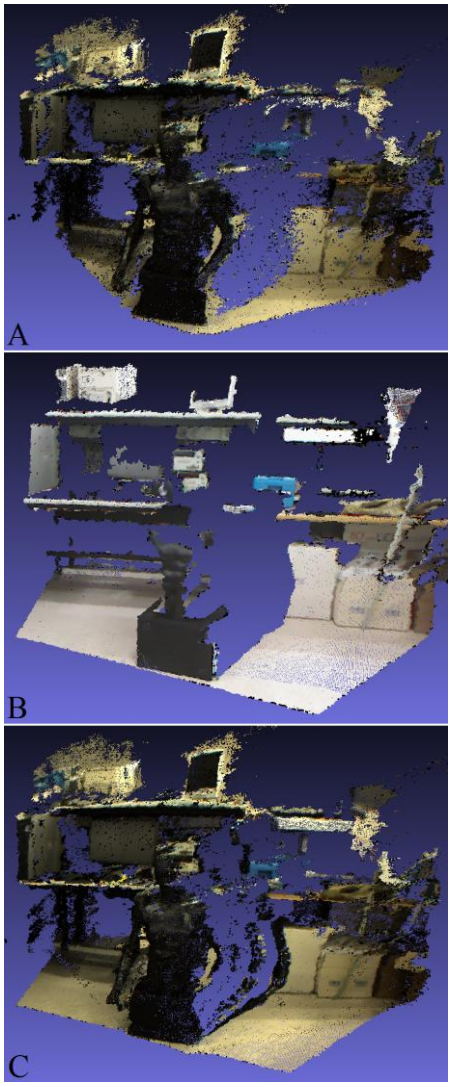

Fig. 2 A. StereFusion. B. KinectFusion, specular surfaces couldn't be reconstructed. C. Stereo+KinectFusion.

Depth image based ICP used in KinectFusion is prone to registration failures in the case of large camera motion, as well as poor 3D structure. Visual odometry is used in [15] to switch from ICP based camera motion solution to visual odometry based solution if the two doesn't agree. ICP is initialized close to the global optimum in [19] using visual odometry solution. We are proposing to use stereo based visual odometry [6] motion estimation for initializing ICP [19], as well as replacing the final ICP solution with odometry solution if there is a disagreement [15]. This strategy exploits both good initialization for ICP and robustness of visual odometry. Additionally our system uses stereo visual odometry instead of a single RGB camera [15],[19] which is expected to be much more accurate.

Cyclical buffers and shift procedures are deployed in [11],[13] for continuous extended mapping of the environment. We use visual odometry navigation solution to decide whether KinectFusion volume needs to be reset. If the cumulative change in rotation and translation exceeds a certain threshold, the system is restarted. A point cloud is saved at every reset of the volume. These point clouds are correctly located with respect to the global coordinates because initial pose of KinectFusion framework is set to the pose given by the visual odometry (i.e. global pose). Even though this procedure gives redundant point clouds due to overlapping regions, these are filtered using voxel grid filter during offline processing. Using visual odometry for stitching point clouds also enable multi-session 3D reconstruction, in which the user turns on and off the KinectFusion reconstruction process for disconnected regions of interest.

III. Architecture And Results

A. Kinect+StereoFusion

Even though stereo is expected to generate less complete and noisier depth maps, it is able to function outdoors. Kinect depth images are replaced with stereo depth images (LibElas [20]) in PCL open source KinectFusion framework [21], which is called StereFusion. A sample reconstruction in StereoFusion is given in Fig. 2A. The proposed system is an alternative to RGB image based sparse reconstruction frameworks (i.e. [6]), once the spatially extended reconstruction is made available (Section 3C). The advantage of StereoFusion over other frameworks is its unprecedented 3D model accuracy due to TSDF representation.

The Kinect and stereo depth maps complement each other [22],[23]. Kinect depth image fails for transparent, specular, flat dark surfaces (Fig. 2B), while stereo depth map is incomplete for low texture regions (Fig. 2A). Stereo (I^s) and Kinect (I^k) depth images can be registered and fused once the external stereo calibration is performed (IR camera of the Kinect and one of the RGB cameras on the Bumblebee). In order to fuse the two depth maps at every frame, point cloud of Kinect depth map (C^k) is computed using the Q matrix: $C^k = Q * I^k$. Then this point cloud is transformed to align with the coordinate system of the stereo using transformation matrix (T_k^s) computed in calibration: $C^s = T_k^s * C^k$. The registered depth image is generated by projecting the transformed point cloud on the image plane of stereo: $I^{ks} = P^s * C^s$. This depth image is fused with the stereo depth image using weighted averaging: $I^f = I^{ks} + w * I^s$.

More complex fusion algorithms [22]-[25] are omitted for real-time operation concerns. A sample reconstruction using the fused depth map is given in Fig 2C, which shows a significant improvement over Kinect-only reconstruction (Fig 2B).

B. Stereo Visual Odometry and ICP

ICP registration used in KinectFusion is erratic. The drift for no motion scenario is shown Fig 3, where visual odometry (red line) shows much more robust behaviour. For utilizing this robustness, we initialized ICP with the solution of visual odometry to avoid local minima and still used visual odometry solution if the final ICP solution deviates significantly from visual odometry.

C. Continuous and Multi-Session KinectFusion

Visual odometry solution is computed continuously in our system. Once the KinectFusion thread is turned on, the pose of the Kinect camera is initialized with the current visual odometry pose, i.e. global pose. The reconstruction is continuous for large area 3D modelling, unless the user turns it off, after which a point cloud is saved. Continuity of the

reconstruction for large areas is achieved by constant monitoring of cumulative camera motion, and once a reset is needed, saving the point cloud and restarting the KinectFusion thread automatically. This procedure produces multiple overlapping point clouds that are correctly located with respect to each other (Fig 4). A very disjoint location can be reconstructed in a separate session, and it can be correctly located in the global coordinates since visual odometry is constantly computing the pose of the camera during operation.

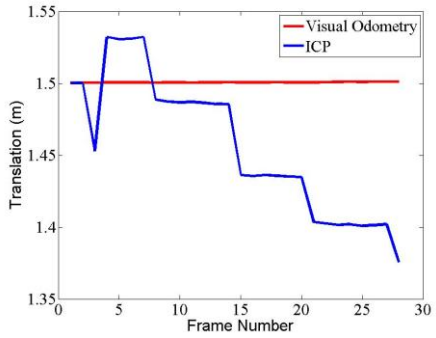

Fig. 3 The drift in translation in one axis during KinectFusion, and the corresponding visual odometry output.

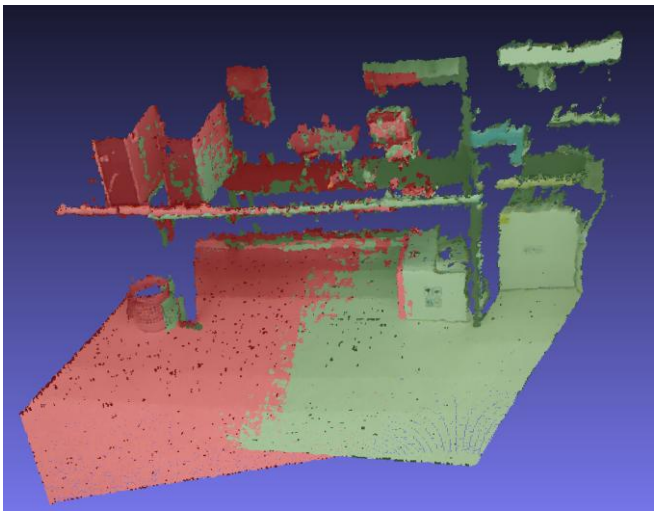

Fig. 4 For close up view of continuous KinectFusion framework, two sessions of reconstruction results are shown in different colors. Small registration error is due to visual odometry correction.

ACKNOWLEDGMENT

This study is supported by HYPERION (FP7) project. We thank Sait Kubilay Pakin for his contribution to the overall system design.

REFERENCES

[1] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *Proceedings Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan, November 2007.

[2] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in Computer Vision (ICCV), 2011 IEEE Int. Conf. on, pp. 2320 –2327, November 2011.

[3] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time Dense Surface Mapping and Tracking," in Proc. of the 2011 10th IEEE Int. Symposium on Mixed and Augmented Reality, ISMAR '11, (Washington, DC, USA), pp. 127–136, 2011.

[4] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Visual odometry and mapping for autonomous flight using an RGB-D camera," in Int. Symposium on Robotics Research (ISRR), (Flagstaff, Arizona, USA), August 2011.

[5] K. Pirker, M. R¨uther, G. Schweighofer, and H. Bischof, "GPSlam: Marrying sparse geometric and dense probabilistic visual mapping," in Proc. of the British Machine Vision Conf., pp. 115.1–115.12, 2011.

[6] A. Geiger, J. Ziegler, and C. Stiller. StereoScan: Dense 3d reconstruction in real-time. In IV, 2011.

[7] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the RGB-D SLAM system," in Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), (St. Paul, MA, USA), May 2012.

[8] C.G. Harris and J.M. Pike, "3D Positional Integration from Image Sequences," Proc. Third Alvey Vision Conf., pp. 233-236, 1987.

[9] C. Tomasi and T. Kanade, "Shape and Motion From Image Streams Under Orthography: A Factorization Method," *Int'l J. Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[10] F. Steinbruecker, J. Sturm, and D. Cremers, "Real-Time Visual Odometry from Dense RGB-D Images," in Workshop on Live Dense Reconstruction with Moving Cameras at the Int. Conf. on Computer Vision (ICCV), November 2011.

[11] T. Whelan, J. McDonald, M. Kaess, M. Fallon, H. Johannsson, and J. Leonard, "Kintinuous: Spatially Extended KinectFusion," in 3rd RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras, (Sydney, Australia), July 2012.

[12] "ReconstructMe FAQ." http://reconstructme.net/usage/#multiscan, August 10th 2012.

[13] "KinectFusion extensions to large scale environments." http://www.pointclouds.org/blog/srcs/fheredia/index.php, August 10th 2012.

[14] H. Roth and M. Vona, "Moving volume KinectFusion," in British Machine Vision Conf. (BMVC), (Surrey, UK), September 2012.

[15] T. Whelan, H. Johannsson, M. Kaess, J.J. Leonard, and J.B. McDonald. Robust real-time visual odometry for dense RGB-D mapping. In IEEE Intl. Conf. on Robotics and Automation, ICRA, Karlsruhe, Germany, May 2013.

[16] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," PAMI, vol. 29, no. 6, pp. 1052–1067, 2007.

[17] C. Audras, A. I. Comport, M. Meilland, and P. Rives, "Real-time dense RGB-D localisation and mapping," in Australian Conf. on Robotics and Automation, (Monash University, Australia), December 2011.

[18] F. Steinbruecker, J. Sturm, and D. Cremers, "Real-Time Visual Odometry from Dense RGB-D Images," in Workshop on Live Dense Reconstruction with Moving Cameras at the Int. Conf. on Computer Vision (ICCV), November 2011.

[19] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments," The Int. Journal of Robotics Research, 2012.

[20] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In ACCV, 2010.

[21] R. B. Rusu and S. Cousins. 3D is here: Point cloud library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 2011.

[22] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, B. Wilburn, and H. Yu, "Edge-preserving photometric stereo via depth fusion," In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[23] W. C. Chiu, U. Blanke, and M. Fritz. Improving the kinect by cross-modal stereo. In BMVC, 2011.

[24] Y. M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Micusik, and S. Thrun. Multi-view image and tof sensor fusion for dense 3d reconstruction. In Proc. of 3DIM 2009, 2009

[25] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In CVPR, 2008.

[26] http://www.filmtools.com/rps-video-stabilizer-with-padded-shoulder-dl-0370.html