TÜBİTAK

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2016) 24: 2756 – 2770
© TÜBİTAK
doi:10.3906/elk-1401-273

Stereo and KinectFusion for continuous 3D reconstruction and visual odometry

Özgür YILMAZ[1,*], **Fatih KARAKUŞ**[2]
[1]Department of Computer Engineering, Turgut Özal University, Ankara, Turkey
[2]Aselsan Inc., MGEO Division, Ankara, Turkey

Received: 30.01.2014 • **Accepted/Published Online:** 02.12.2014 • **Final Version:** 15.04.2016

Abstract: Robust and accurate 3D reconstruction of a scene is essential for many robotic and computer vision applications. Although recent studies propose accurate reconstruction algorithms, they are only suitable for indoor operation. We are proposing a system solution that can accurately reconstruct the scene both indoors and outdoors, in real time. The system utilizes both active and passive visual sensors in conjunction with peripheral hardware for communication and suggests an accuracy improvement in both reconstruction and pose estimation accuracy over state-of-the-art SLAM algorithms via stereo visual odometry integration. We also introduce the concept of multisession reconstruction, which is relevant for many real-world applications. In our solution to this concept, distinct regions in a scene can be reconstructed in detail in separate sessions using the KinectFusion framework and merged into a global scene using continuous visual odometry camera tracking.

Key words: 3D reconstruction, SLAM, stereo, Kinect, visual odometry, iterative closest point, fusion

1. Introduction

Understanding the geometry of a scene is essential for artificial vision systems. Accurate 3D modeling of specific objects or a scene can also be very beneficial for archival or urban planning problems. Due to the richness of their output, 3D reconstruction algorithms are widely used in many computer vision applications (Figure 1).

Many robust solutions have been proposed in 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 (Figure 2). 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, the KinectFusion algorithm [3] was a huge step towards real-time operating 3D reconstruction systems given its reconstruction accuracy and speed (Figure 3). It uses a volumetric representation of the scene called truncated sign distance function (TSDF) and the fast iterative closest point (ICP) algorithm for camera motion estimation. TSDF is a volumetric representation of the scene, where a fixed-size 3D volume (e.g., a cube of 3 m in size) of a fixed spatial resolution is initialized. This volume is

[*]Correspondence: ozyilmaz@turgutozal.edu.tr

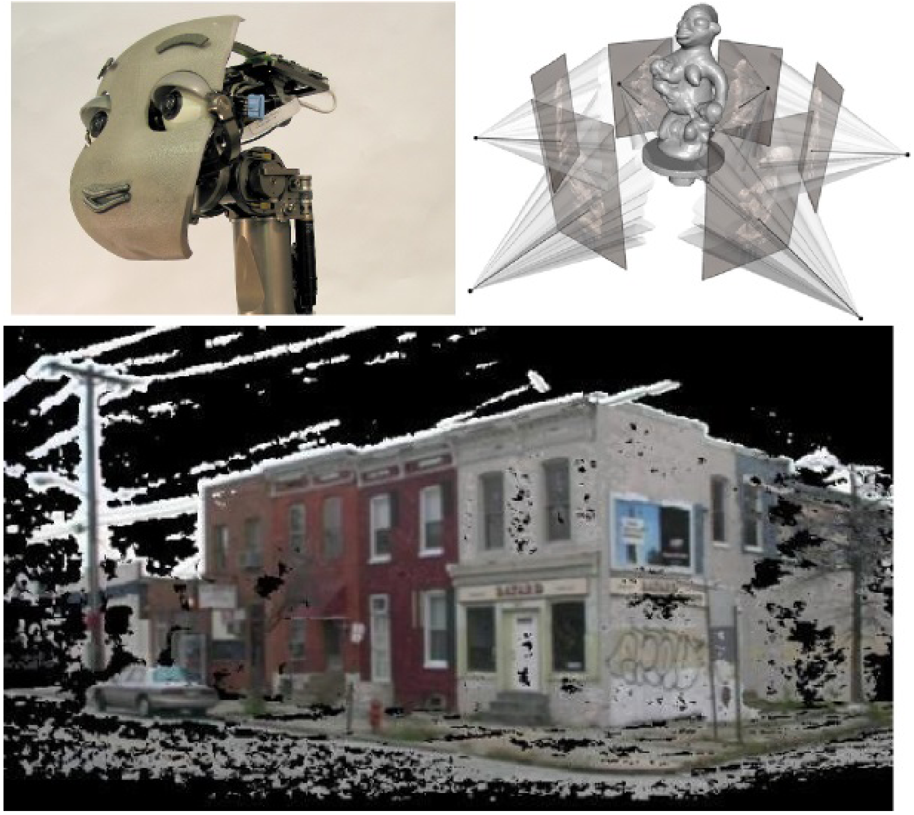


Figure 1. Three possible applications of 3D reconstruction. In robot vision (upper left), detailed geometric inferences are performed in the generated 3D models. 3D reconstruction is essential in archiving the models of archaeological artifacts (upper right). In urban modeling, 3D models of the building or a larger region are extracted for planning and archiving purposes (image resources from people.csail.mit.edu, carlos-hernandez.org, and www.cs.unc.edu/~marc are used for this figure).

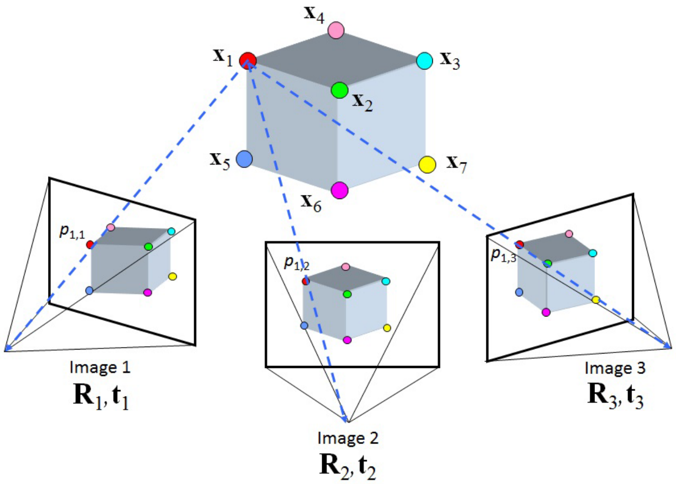


Figure 2. Structure from Motion (SfM) process is illustrated. The structure in the world (cube) is imaged from multiple viewpoints (Image 1, Image 2, Image 3). By tracking (see dashed lines) the pixel locations ($p_{1,1}$, $p_{1,2}$, ...) of specific features (x_1, x_2, ...) in the images, both the camera motion (R_1, T_1; R_2, T_2; ...) and the 3D model of the structure are estimated. Adapted from Svetlena Lazebnik's Computer Vision course material.

divided into equally sized voxels (pixels in three dimensions), and the distance value to the closest surface is kept in each voxel, being updated at every frame as new range measurements are acquired. TSDF has many advantages over other representations such as meshes, the most important of which is its ability to efficiently deal with multiple measurements of the same surface. Its major disadvantage is its high memory requirement. 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 in 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 the scene is poor in 3D structure (i.e. flat regions). Voxel-based scene representation is problematic for reconstruction of a large area due to memory limitations (Figure 4). In recent studies, the KinectFusion algorithm was modified and extended to include solutions to these shortcomings.

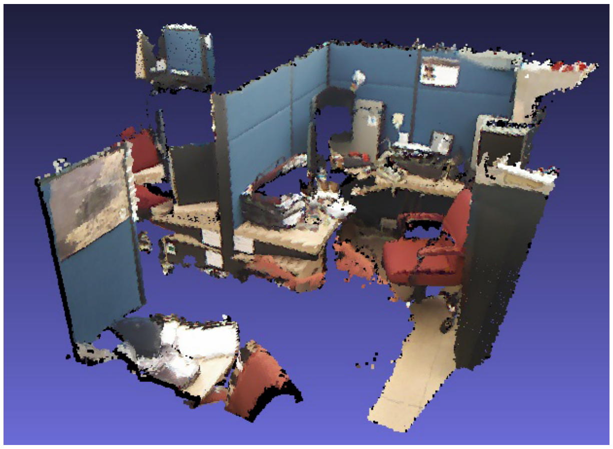


Figure 3. A sample reconstruction of the scene with the KinectFusion algorithm using the Kinect sensor. It provides a detailed 3D model of an indoor scene, thanks to its TSDF volume representation.

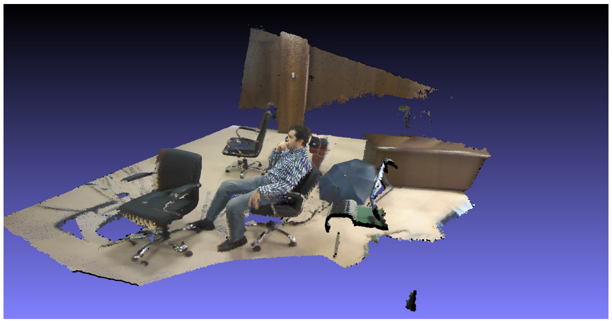


Figure 4. KinectFusion algorithm's limit for the reconstruction size is demonstrated. In the image there is a sharp slanted edge on the left side, which is the border of the allowable volume for reconstruction. Even though the Kinect sensor gathers data to the left of this border, it cannot be included in the 3D model due to memory limitations.

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] (see also: KinectFusion extensions to large-scale environments, http://www.pointclouds.org/blog/srcs/fheredia/index.php, 10 August 2012). The proposed approaches are mainly based on automatically detecting if the camera moves out of the defined volume and reinitiating the algorithm, after saving the previously reconstructed volume as TSDF [14] or saving into a more efficient 3D model 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. Also, the odometry algorithm they used to aid ICP registration was RGB-D camera-based, which is expected to be inferior to stereo odometry approaches. We propose a system that uses the best of both worlds: indoor/outdoor operation with fused sensor data and stereo odometry for accurate registration.

Additionally, the concept that we build is different from the ones studied in the recent literature: robust multisession 3D reconstruction for both indoor and outdoor operation. For some applications, due to limited energy resources, memory capacity, or time limitations (e.g., HYPERION Project, FP7), there is no need to reconstruct a very large region as a whole but rather some disconnected areas of interest in the region, executing reconstruction in multiple sessions (Figure 5). The 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, the 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:

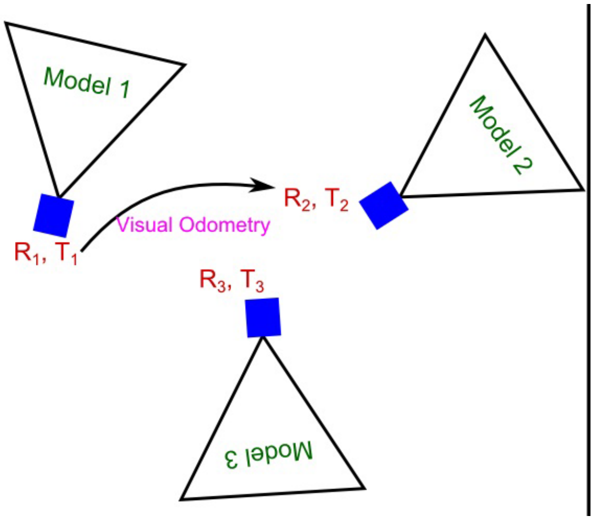


Figure 5. Multisession 3D reconstruction. The 3D models of distinct regions are reconstructed using the KinectFusion framework and these models are relatively located (R, rotation, T, translation) on a predefined coordinate system by tracking the camera at all times using stereo-based visual odometry.

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 the 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 (Figure 5).

Although we are providing theoretical novelty such as multisession reconstruction, the main contributions of our study are practical, aiming at building a robust reconstruction system with well-known existing algorithms. The system that we propose has obvious practical advantages over the existing ones. One major advantage is outdoor operation, and another one is its improved accuracy due to stereo odometry. We think these are very important enhancements to 3D reconstruction systems.

The main contributions of our study are:

- introduction of stereo in the KinectFusion framework,
- utilization of stereo visual odometry for improving registration and global localization of separate 3D models,
- design of a multisession 3D reconstruction concept,
- and a complete system solution to 3D reconstruction.

2. System and algorithmic approach

2.1. 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 (Figure 6A) for imaging, a padded shoulder image stabilizer (Figure 6B) 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 12 V battery power supply) and stereo images even in rough terrains. The images are uploaded to a workstation through wireless transfer and processed.

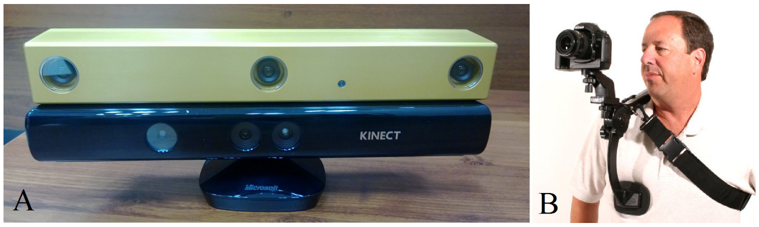


Figure 6. The system components. A) Kinect plus Bumblebee XB3 stereo rig used in the system. B) Padded shoulder image stabilizer for image acquisition. Brand name: RPS. These components along with a notebook PC allow for fine 3D reconstruction of outdoor scenes in rough terrains.

2.2. Algorithmic approach

Stereo is an essential aid to Kinect for outdoor operation since Kinect is not able to give depth maps under sunlight 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 the KinectFusion framework. We adopt a simple and fast approach for fusing two depth maps that works quite well: weighted averaging of pixels after registration.

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 was used in [15] to switch from ICP-based camera motion solution to visual odometry-based solution if the two do not agree. ICP was initialized close to the global optimum in [19] using visual odometry solution. The aforementioned odometry solutions are based on RGB-D camera; however, 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 an odometry solution if there is a disagreement [15]. This strategy exploits both good initialization for ICP and robustness of stereo visual odometry. Additionally, our system uses stereo visual odometry instead of a single RGB-D camera [15,19], which is expected to be more accurate [20]. The improved accuracy of stereo odometry compared to monocular is due to:

1. Only two views are enough for feature matching in stereo whereas monocular requires three views,
2. 3D structure is computed in a single shot in stereo, but adjacent frames are used in monocular, which introduce larger positional drifts, especially for small camera motions.

Mono and stereo visual odometry options are not compared in our study since stereo is widely accepted to be superior and this comparison is out of the scope of this paper.

In addition to the original problem of multisession 3D reconstruction (Figure 5), it is possible to build the 3D model of a large environment continuously (Figure 7). Cyclical buffers and shift procedures are deployed in [11] for continuous extended mapping of the environment. We use a visual odometry navigation solution to decide whether the 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 the initial pose of the 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 a voxel grid filter during offline processing. We should note that using visual odometry for stitching point clouds also enables multisession 3D reconstruction, in which the user turns on and off the StereoKinectFusion reconstruction process for disconnected regions of interest.

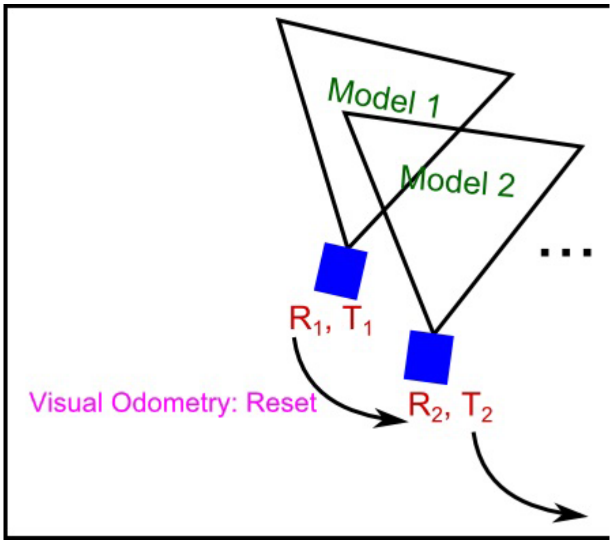
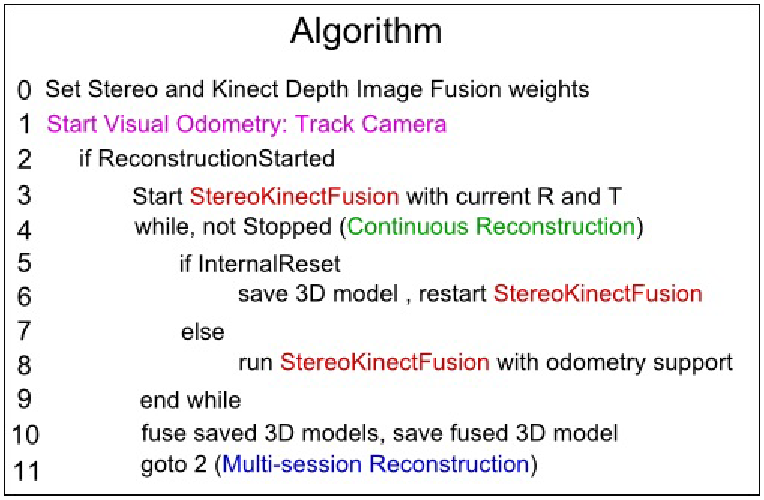


Figure 7. Continuous 3D reconstruction of a large scene is also possible within our framework. Visual odometry decides whether there is a need for reset; if yes, the previous model is saved as a point cloud and reconstruction is restarted. Since the initial poses (e.g., R_1, T_1, R_2, T_2) of the models are known from visual odometry, a global 3D model can be built by merging the individual models.

The algorithm is given in a box (Figure 8) and colored for emphasizing important subroutines. The most important initialization from the user is the stereo and Kinect depth map fusion weights, set according to the lighting conditions of the environment (Line 0). Stereo visual odometry starts right away and keeps running during the execution as the background process (Line 1). If the user does not press the Reconstruct button, the StereoKinectFusion process is not initiated and visual odometry keeps tracking the camera pose in the background. However, as soon as the button is pressed, the reconstruction subroutine is called (Line 3) and it is initiated with the current pose of the camera (R and T) for global registration of multisession reconstructions (see Figure 5). The reconstruction keeps building the model as long as the user does not press the Stop button (Line 4), and if the volume limits of the StereoKinectFusion process are reached (InternalReset, Line 5), an intermediate 3D model is saved as the point cloud and the process is restarted with fresh memory (Line 6). Remember that the camera pose estimation from stereo odometry aids the ICP solution during StereoKinectFusion (Line 8). If reconstruction is stopped by the user, the 3D models saved during continuous reconstruction (while loop, see Figure 7) are fused to get a large and complete 3D model of the scene (Line 10). Once the fused model is saved as a point cloud file, the system is ready for another session of reconstruction and goes to Line 2, waiting for a Start command from the user (different models in Figure 5).



```
Algorithm

0  Set Stereo and Kinect Depth Image Fusion weights
1  Start Visual Odometry: Track Camera
2      if ReconstructionStarted
3              Start StereoKinectFusion with current R and T
4              while, not Stopped (Continuous Reconstruction)
5                  if InternalReset
6                          save 3D model , restart StereoKinectFusion
7                  else
8                          run StereoKinectFusion with odometry support
9              end while
10             fuse saved 3D models, save fused 3D model
11             goto 2 (Multi-session Reconstruction)
```

Figure 8. Algorithm pseudocode. It is a high level description of the algorithm, yet this is enough for replication purposes.

3. Experiments and results

3.1. 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 [21]) in the PCL open source KinectFusion framework [22], which is called StereoFusion. A sample reconstruction in StereoFusion is given in Figure 9A. 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 3.3). The advantage of StereoFusion over other frameworks is its high 3D model accuracy due to TSDF model representation.

The Kinect and stereo depth maps complement each other [23,24]. The Kinect depth image fails for transparent, specular, flat dark surfaces (Figure 9B), while the stereo depth map is incomplete for low texture regions (Figure 9A). In the scene of Figure 9, there is no transparent surface, but the two computer monitors are flat dark surfaces and the mannequin has specularities. Kinect reconstruction fails to capture these regions. Stereo (I^s) and Kinect (I^k) depth images can be registered and fused once the external stereo calibration is performed (IR camera of Kinect and one of the RGB cameras on Bumblebee).

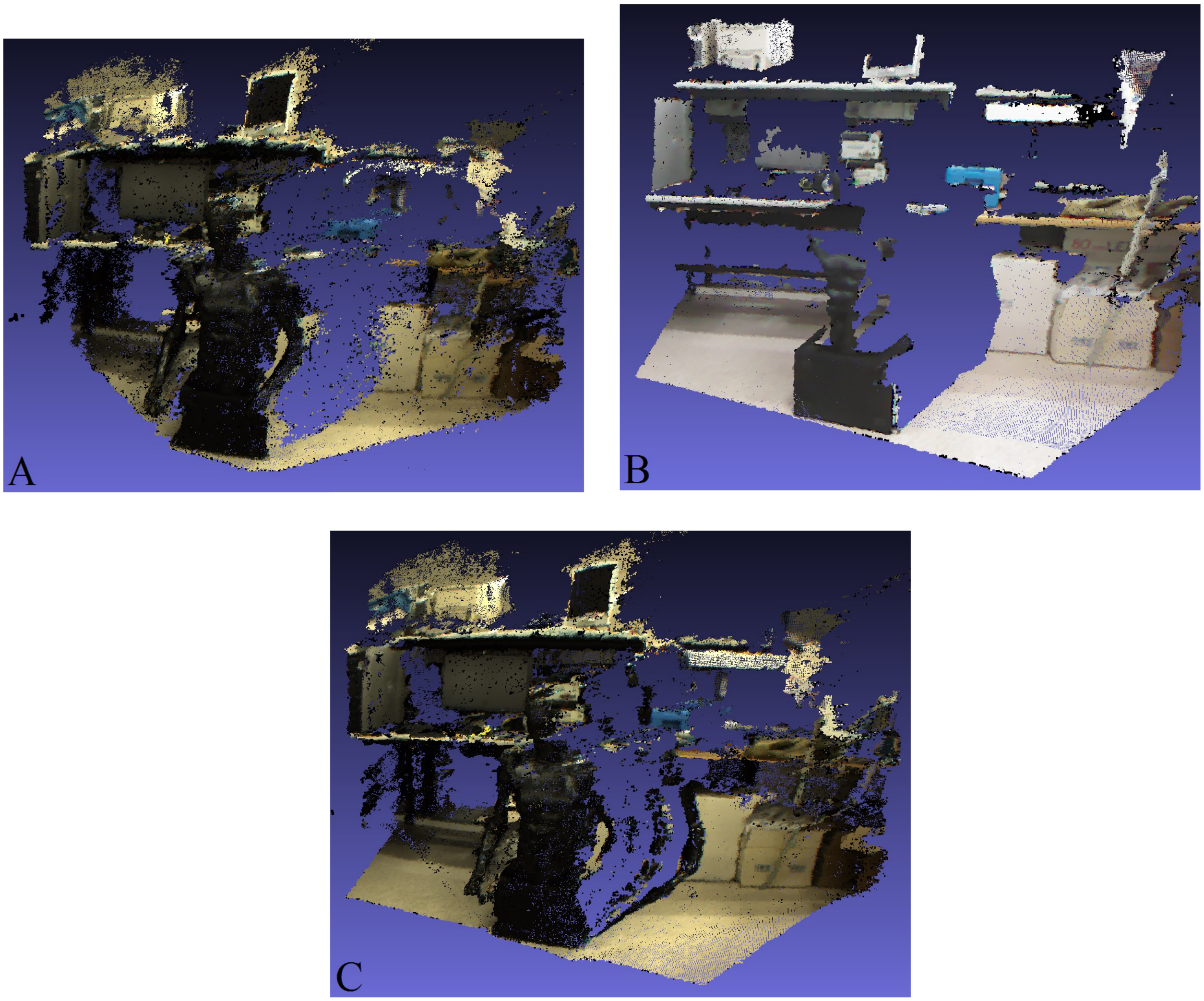


Figure 9. A) StereoFusion, reconstruction using only stereo depth map. B) KinectFusion: reconstruction using only Kinect depth map, specular surfaces could not be reconstructed. C) Stereo+KinectFusion: merging stereo and Kinect depth maps for improved reconstruction. In the scene there is no transparent surface, but the two computer monitors are flat dark surfaces and the mannequin has specularities. Kinect reconstruction fails to capture these regions.

The calibration procedure computes external parameters such as:

1. Q matrix, which generates the point cloud from the depth map.
2. The transformation matrix that rotates and translates the point cloud of the Kinect sensor to align with the point cloud of stereo.
3. The projection matrix that generates the synthetic depth map of Kinect on the same coordinate system with stereo depth map, ready to be used for fusion of the two depth maps.

In order to fuse the two depth maps at every frame:

1. The point cloud of the Kinect depth map (C^k
2. is computed using the Q matrix: $C^k = Q \ times \ I^k$.
3. Then this point cloud is transformed to align with the coordinate system of the stereo using the transformation matrix (T_k^s) computed in calibration: $C^s = T_k^s \ times \ C^k$.
4. The registered depth image is generated by projecting (projection matrix computed using pose estimation solution) the transformed point cloud on the image plane of stereo: $I^{ks} = P^s \ times \ C^s$.
5. This depth image is fused with the stereo depth image using weighted averaging: $I^f = I^{ks} + \ w \ times \ I^s$. The weight gets discrete values: 0 (Kinect only), 0.5 (balanced averaging), and 1 (stereo only) (see Figure 9).

More complex fusion algorithms [23–27] are omitted for real-time operation concerns. A sample reconstruction using the fused depth map is given in Figure 9C, which shows a significant improvement over Kinect-only reconstruction (Figure 9B) due to specularities and dark surfaces. Calibration between Kinect and stereo cameras is essential and it is important to report this procedure for experimental replication purposes. The main problem in calibration is estimation of the external calibration parameters, i.e. relative location and orientation of the two sensors. In order to compute these, a checkerboard pattern is used along with a standard calibration toolbox. The infrared image from the Kinect (IR projector turned off via duct tape) and the stereo images from Bumblebee are simultaneously captured. Stereo calibration is performed on these two images: the left image of the stereo and the IR image of the Kinect. Since the calibration procedure estimates the external calibration parameters, we are able to extract the relative location and orientation between the two sensors.

3.2. Outdoor reconstruction

We used our system outside on a sunny day for reconstruction. By setting the weight of the Kinect depth map to zero, we tested StereoFusion (see also Figure 9A). The results show that StereoFusion gives high quality and detailed 3D reconstruction (Figure 10) compared to other alternatives (e.g., [6] or [28]), due to TSDF representation and small registration error of stereo odometry. TSDF representation is capable of accurately merging multiple measurements of the same surface that are acquired as the camera moves and filling in the holes as new measurements are added [3]. This capability of the KinectFusion framework is a very powerful feature and it is aided with a better registration strategy in our system, which results in detailed and accurate 3D models.

For outdoor reconstruction, we are exploiting the best options for both the model representation and registration, but with a cost: real-time (15 fps) performance can only be achieved on a powerful computer (16 CPU cores and 1536 GPU cores). There are various reasons for high computational power demands:

1. The standard KinectFusion algorithm is already computationally demanding mainly due to a volumetric integration subroutine that manipulates TSDF volume at every frame. It utilizes both CPU and GPU, but main horsepower is provided by GPU.
2. We have added stereo depth map generation and depth map fusion that runs on CPU.
3. We have added CPU threads for stereo visual odometry.

Overall, the significant additions to the standard KinectFusion framework are stereo depth map and stereo odometry blocks. The computation time of these algorithms can be found in [21] and [9].

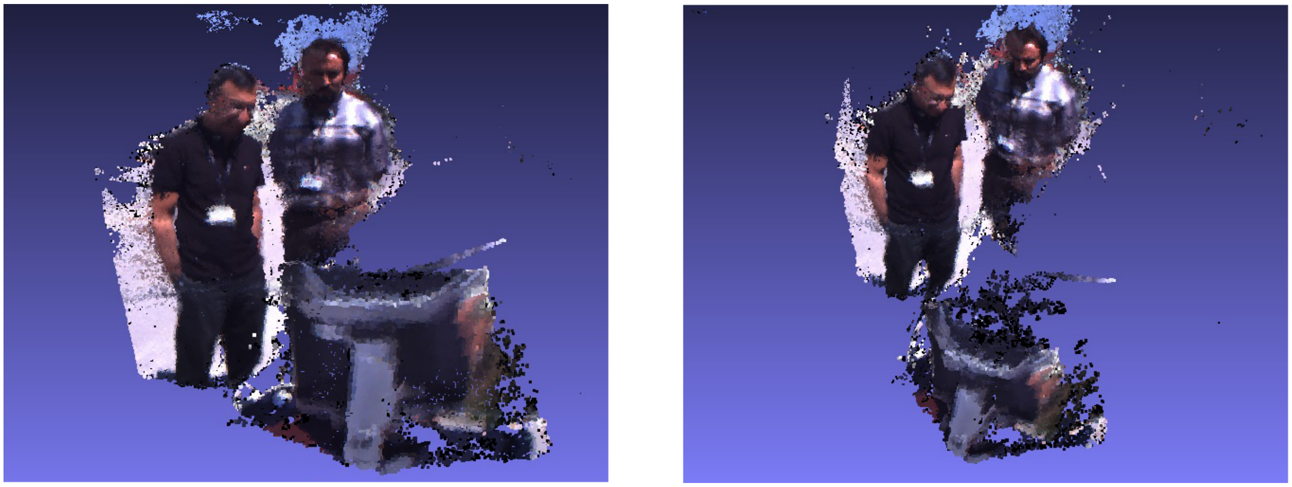


Figure 10. Outdoor reconstruction of the StereoFusion algorithm. The 3D model is accurate and detailed.

3.3. Stereo visual odometry and ICP

ICP registration used in KinectFusion is erratic. The drift for no motion scenario is shown in Figure 11A, where visual odometry (red line) [6] shows much more robust behavior. In Figure 11, the camera is standing still but the ICP solution is drifting, which shows the instability of the ICP algorithm for odometry solutions. The scene is identical to the one given in Figure 9 and it has both planar regions and convoluted 3D surfaces. In order to utilize the robustness of stereo odometry, we initialized ICP with the solution of visual odometry to avoid local minima and still used the visual odometry solution if the final ICP solution deviated significantly (0.03 m) from visual odometry, as in [15]. For stereo odometry, we are using the code generously provided by the authors of the algorithm [6].

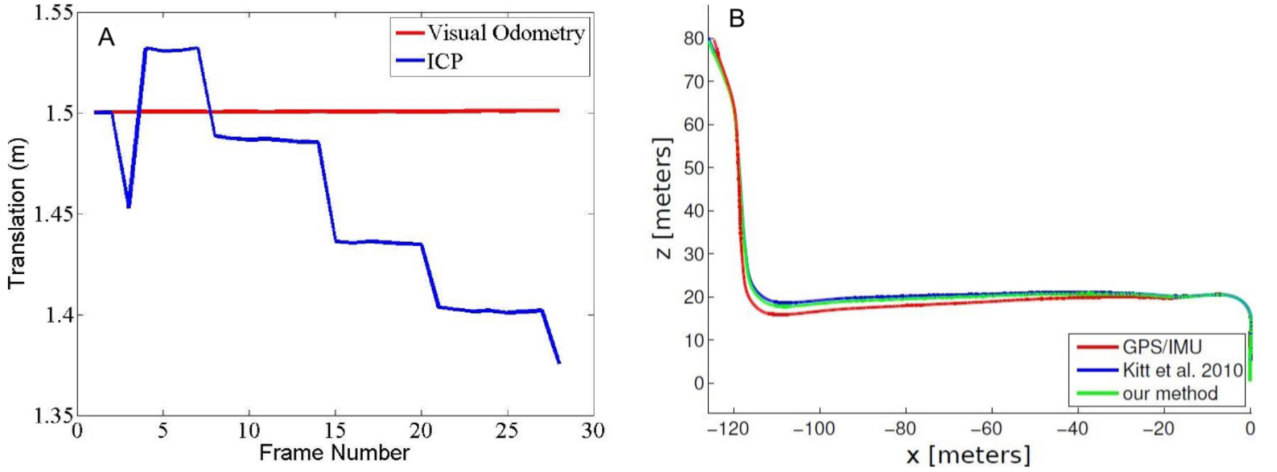


Figure 11. A) The drift in translation in one axis during KinectFusion (without stereo odometry support), and the corresponding visual odometry output (algorithm from [6]). The camera is standing still but the ICP solution is drifting, which shows the instability of the ICP algorithm for an odometry solution. B) Visual odometry error computed on a standard dataset (adapted from [6]). Red curve is the ground truth computed by GPS/IMU navigation solution. Blue curve is another odometry algorithm's [29] error, and green is the algorithm we adopted for visual odometry.

The amount of improvement in positional drift due to stereo odometry is more visible when reconstructing a large area (Figure 11B, adapted from [6]; the utilized algorithm is compared with another algorithm from [29]). In large area reconstruction, the overall 3D model of the scene becomes significantly distorted in time due to the accumulation of positional error. In general, the effect of this drift is not visible in the standard KinectFusion application [3] because the size of the model is limited (a cube of 3 m size) and the maximum amount of drift that can be experienced is considerably small. Therefore, positional error and 3D model deficiencies due to low quality registration (i.e. ICP) are not visible unless large area reconstruction is attempted. The contribution of our approach is introduction of a very accurate odometry module that enhances both positional estimates and the produced 3D model, especially for large area reconstruction. Hence, the improvement in accuracy of visual camera tracking (Figure 5) is most valuable when considered with multisession and large scale reconstructions (next section).

3.4. Continuous and multisession KinectFusion

The 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

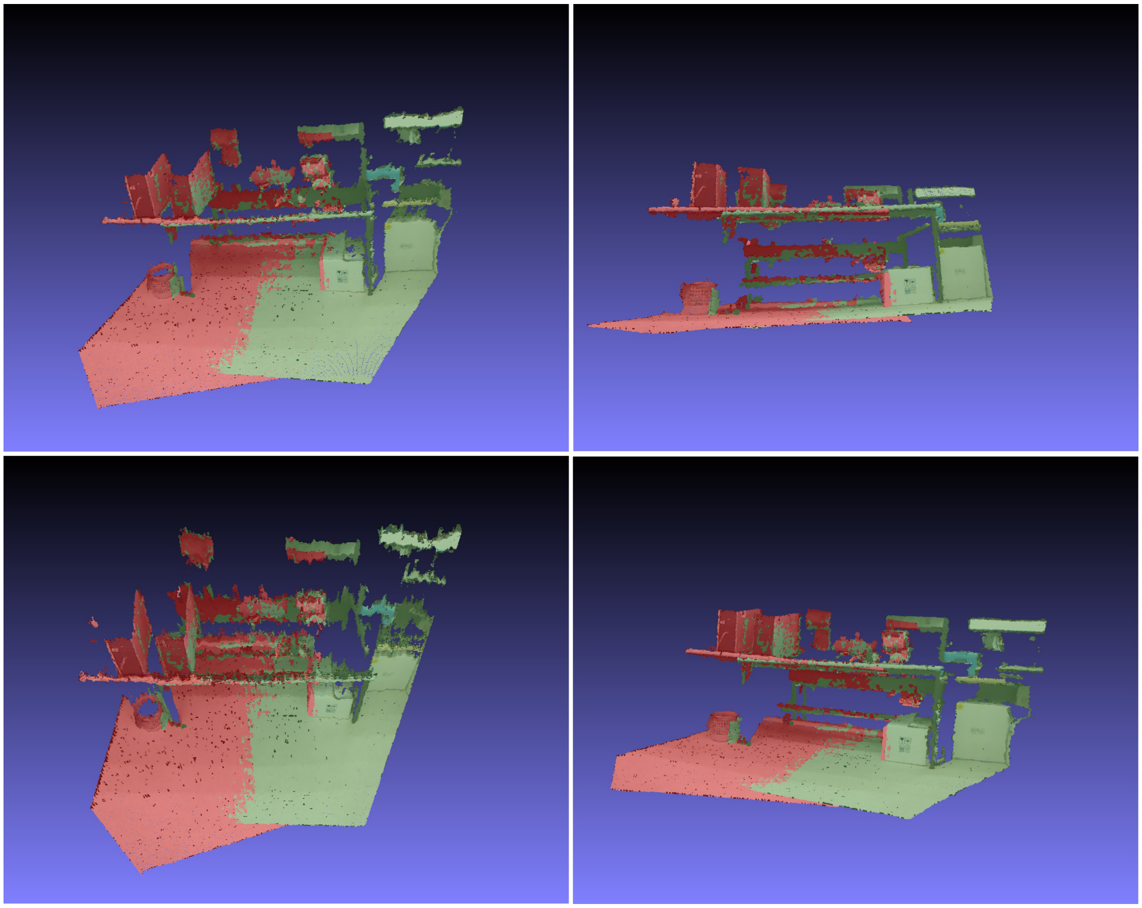


Figure 12. For close-up view of the continuous KinectFusion framework, two sessions of reconstruction results are shown in different colors. The two 3D models are perfectly aligned. Small registration error is due to visual odometry correction.

reconstruction is continuous for large area 3D modeling, 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. The reset is initiated once the camera is 1.5 m away from the cubic volume center. This procedure produces multiple overlapping point clouds that are correctly located with respect to each other. A very disjoint location can be reconstructed in a separate session (Figure 7), and it can be correctly located in the global coordinates since visual odometry is constantly computing the pose of the camera during operation. The qualitative results (Figure 12) show that the system is able to locate separate sessions very accurately. Multisession reconstruction utilizes the global pose estimate provided by the stereo visual odometry module. Thus, relative registration between separate models is only affected by the drift in stereo odometry. It is the sole source of relative registration errors. A very accurate algorithm is implemented in our system, which is illustrated in Figure 11B. Therefore, we are using a state-of-the-art odometry algorithm that enables accurate multisession 3D reconstruction.

In order to illustrate the improved accuracy of registration due to stereo odometry, a whole room was reconstructed in multiple sessions enforcing a closed loop, and for the registration subroutine only ICP (original KinectFusion) or the proposed algorithm (stereo odometry-aided ICP) was used. Figure 13 compares the registration accuracy of only ICP (Figure 13a) and the proposed approach (Figure 13b) by using the ground plane as the anchor. It is observed that the misalignment on the ground is larger in the ICP approach. Figure 14 shows shadows created by misregistration of multiple sessions in the ICP case, while the proposed approach did not produce such an artifact, although the reconstruction seems more blurry. In Figure 15, again a failure in the ICP case is illustrated, in which the wastebasket and the computer monitors (black objects next to the wastebasket) were not successfully reconstructed.

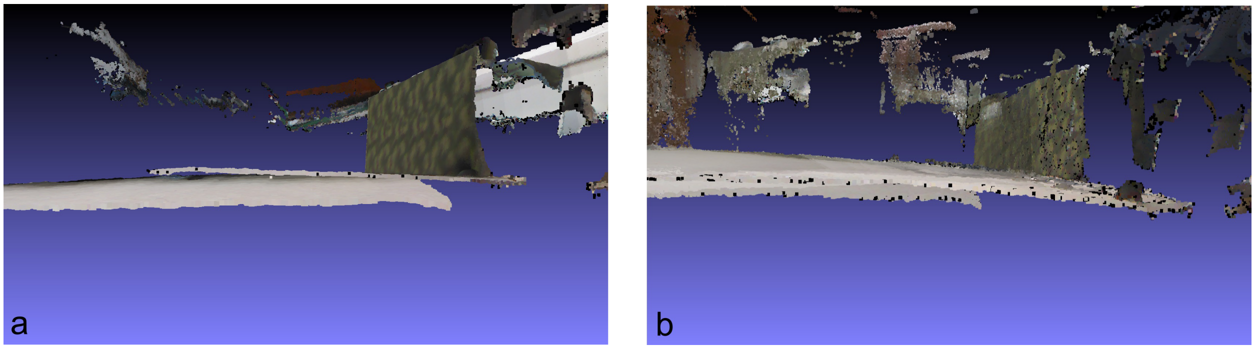


Figure 13. The registration comparison of ICP (a) and the proposed approach (b) using the ground plane. The reconstruction of a room was executed in multiple sessions, and the loop was closed. It is observed that the ground planes that are reconstructed in separate sessions do not overlap well in the ICP approach (a).

3.5. Quantitative analysis and dataset

Our system uses a very rich image acquisition setup: stereo and Kinect. Unfortunately there is no standard annotated dataset for quantitative analysis of our system, so we have to create our own dataset. However, a dataset setup for our purposes requires very expensive equipment (motion capture for registration accuracy and laser scanner for reconstruction accuracy). Even though we are reporting the registration robustness (Figure 11A) and 3D reconstruction quality (Figures 9, 10, 12, 13, and 14) of our system, there are more detailed experiments that need to be performed. This is the future work of our study.

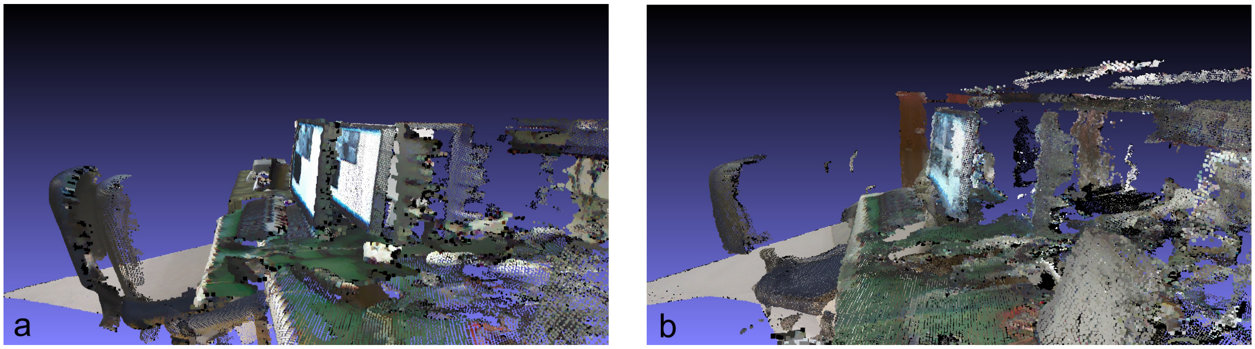


Figure 14. The registration comparison of ICP (a) and the proposed approach (b). Shadows are observed in the ICP approach due to large registration errors, which is not observed in the proposed approach.

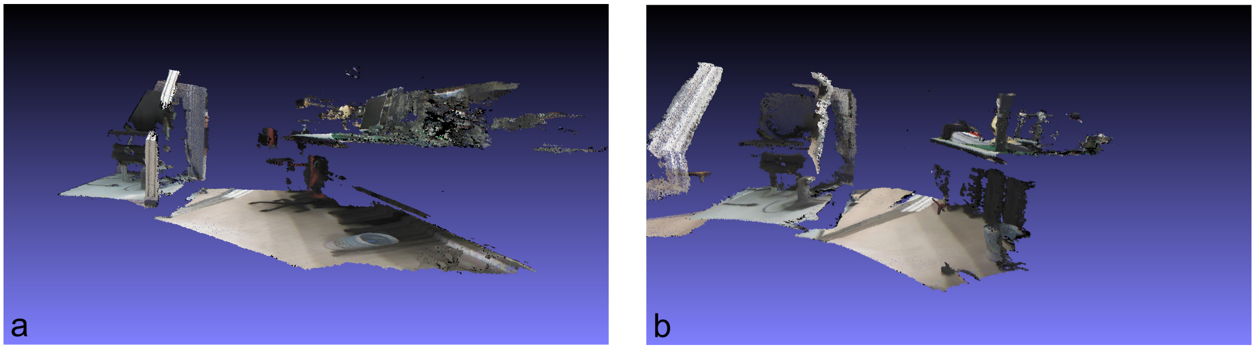


Figure 15. The reconstruction failure of the ICP-based algorithm (a) is illustrated, and the reconstruction of the same scene with the proposed approach is given (b). In (a), the wastebasket and the monitors (black objects next to the wastebasket) are not reconstructed successfully, while the proposed approach (b) does not show these artifacts.

4. Conclusion and future work

We introduced the usage of stereo depth maps into the KinectFusion framework and showed that the outdoor reconstruction is of high quality. Fusion of stereo and Kinect depth maps gives a superior 3D model for indoor reconstruction. Also, a stereo odometry navigation solution is used to aid ICP-based registration in KinectFusion, which improved the robustness and reduced the drift error of the framework. Additionally, we introduced the multisession 3D reconstruction concept, which is useful in many real-life applications. In this concept, distinct regions in a scene can be reconstructed separately in different sessions, and these 3D models can accurately be placed in a global coordinate system using the stereo odometry. As future work we are planning to create a stereo+Kinect dataset, execute more quantitative analyses, and explore the benefits of loop closure techniques in our framework [30].

Acknowledgments

This study was supported by the HYPERION (FP7) project and Turgut Özal University BAP Project Number 006-10-2013, titled “Stereo ve Kinect Kameralar ile 3 Boyutlu Modelleme”. We thank Sait Kubilay Pakin for his contribution to the overall system design.

References

[1] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces. In: Proceedings of the Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality; November 2007; Nara, Japan.

[2] Newcombe RA, Lovegrove SJ, Davison AJ. DTAM: Dense tracking and mapping in real-time. In: International Conference on Computer Vision; November 2011. pp. 2320-2327.

[3] Newcombe RA, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison AJ, Kohli P, Shotton J, Hodges S, Fitzgibbon A. KinectFusion: Real-time dense surface mapping and tracking. In: Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality; 2011; Washington, DC, USA. pp. 127-136.

[4] Huang AS, Bachrach A, Henry P, Krainin M, Maturana D, Fox D, Roy N. Visual odometry and mapping for autonomous flight using an RGB-D camera. In: International Symposium on Robotics Research; August 2011; Flagstaff, AZ, USA.

[5] Pirker K, Rüther M, Schweighofer G, Bischof H. GPSlam: Marrying sparse geometric and dense probabilistic visual mapping. In: Proceedings of the British Machine Vision Conference; 2011. pp. 115.1-115.12.

[6] Geiger A, Ziegler J, Stiller C. StereoScan: Dense 3d reconstruction in real-time. In: Intelligent Vehicles Symposium; 2011.

[7] Endres F, Hess J, Engelhard N, Sturm J, Cremers D, Burgard W. An evaluation of the RGB-D SLAM system. In: Proceedings of the IEEE International Conference on Robotics and Automation; May 2012; St. Paul, MN, USA.

[8] Harris CG, Pike JM. 3D positional integration from image sequences. In: Proceedings of the Third Alvey Vision Conference; 1987. pp. 233-236.

[9] Tomasi C, Kanade T. Shape and motion from image streams under orthography: a factorization method. Int J Comput Vis 1992; 9: 137-154.

[10] Steinbruecker F, Sturm J, Cremers D. Real-time visual odometry from dense RGB-D images. In: Workshop on Live Dense Reconstruction with Moving Cameras at the International Conference on Computer Vision; November 2011.

[11] Whelan T, McDonald J, Kaess M, Fallon M, Johannsson H, Leonard J. Kintinuous: spatially extended KinectFusion. In: 3rd RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras; July 2012; Sydney, Australia.

[12] Meilland M, Comport AI. On unifying key-frame and voxel-based dense visual slam at large scales. In: IEEE International Conference on Intelligent Robots and Systems; 2013.

[13] Chen J, Bautembach D, Izadi S. Scalable real-time volumetric surface reconstruction. In: ACM Transactions on Graphics 2013; 32: 113:1-113:8.

[14] Roth H, Vona M. Moving volume KinectFusion. In: British Machine Vision Conference; September 2012; Surrey, UK.

[15] Whelan T, Johannsson H, Kaess M, Leonard JJ, McDonald JB. Robust real-time visual odometry for dense RGB-D mapping. In: IEEE Internatinoal Conference on Robotics and Automation; May 2013; Karlsruhe, Germany.

[16] Davison AJ, Reid ID, Molton ND, Stasse O. Monoslam: Real-time single camera slam. PAMI 2007; 29: 1052-1067.

[17] Audras C, Comport AI, Meilland M, Rives P. Real-time dense RGB-D localisation and mapping. In: Australian Conference on Robotics and Automation; December 2011; Monash University, Australia.

[18] Steinbruecker F, Sturm J, Cremers D. Real-time visual odometry from dense RGB-D images. In: Workshop on Live Dense Reconstruction with Moving Cameras at the International Conference on Computer Vision; November 2011.

[19] Henry P, Krainin M, Herbst E, Ren X, Fox D. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. Int J Robot Res 2012; 31: 647-663.

[20] Scaramuzza D, Fraundorfer F, Visual odometry. IEEE Robot Autom Mag 2011; 18: 80-92.

[21] Geiger A, Roser M, Urtasun R. Ef?cient large-scale stereo matching. In: ACCV; 2010.

[22] Rusu RB, Cousins S. 3D is here: Point cloud library (PCL). In: IEEE International Conference on Robotics and Automation; May 2011; Shanghai, China.

[23] Zhang Q, Ye M, Yang R, Matsushita Y, Wilburn B, Yu H. Edge-preserving photometric stereo via depth fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2012.

[24] Chiu WC, Blanke U, Fritz M. Improving the Kinect by cross-modal stereo. In: BMVC; 2011.

[25] Kim YM, Theobalt C, Diebel J, Kosecka J, Micusik B, and Thrun S. Multi-view image and TOF sensor fusion for dense 3D reconstruction. In: Proceedings of 3DIM; 2009.

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

[27] Wang Y, Jia Y. A fusion framework of stereo vision and Kinect for high quality dense depth maps. In: Proceedings of the 11th International Conference on Computer Vision, ACCV Workshops; 2012. pp. 109-120.

[28] Furukawa, Y, Curless B, Seitz SM, Szeliski R. Towards internet-scale multiview stereo. In: CVPR; 2010.

[29] Kitt B, Geiger A, Lategahn H. Visual odometry based on stereo image sequences with RANSAC- based outlier rejection scheme. In: IV; 2010.

[30] Angeli A, Filliat D, Doncieux S, Meyer JA. Real-time visual loop-closure detection. In: IEEE International Conference on Robotics and Automation; 2008. pp. 1842-1847.