IMU aided RGB-D SLAM

Usman Qayyum, Qaisar Ahsan and Zahid Mahmood
Center of Excellence in Science and Technology
Islamabad, Pakistan
mrusmanqayyum@gmail.com

Abstract

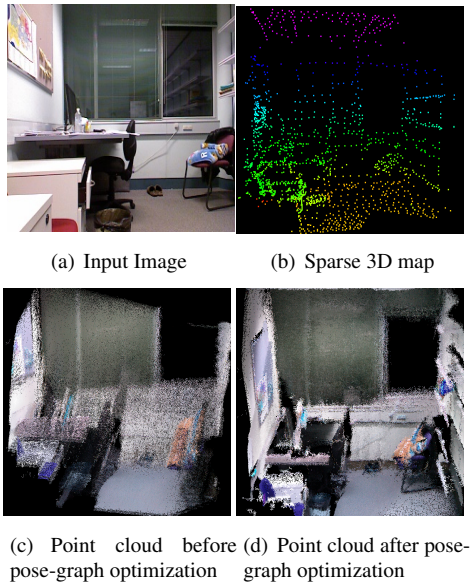
In recent years, a low-cost range sensing technology (such as Kinect) has attracted many researchers in vision community. Despite being promising for indoor navigation, continuous operation in a challenging environments is still a fundamental problem. This paper presents a real-time approach of combining range-color (RGB-D) cues with inertial information in a loosely-coupled framework for Simultaneous Localization and Mapping (SLAM). The method handles the degenerate cases arises in the RGB-D approaches i.e limited range of Kinect (approx. 4m), texture-less/planar environmental ambiguities. Experimental results are provided for indoor lab environment and evaluated against a centimeter-level accurate ground truth data. We showed that the proposed approach significantly reduces error in RGB-D pose estimation by bounding the errors implicitly by the inertial gravity vectors and accelerometer data.

1 Introduction

Real-time localization and mapping for complex indoor environment is a core problem and has a long history in computer-vision research. Various sensor modalities have been used to do SLAM, however vision sensor has got huge success in recent years [1]. A monocular camera based approaches have been proposed by number of researchers, however it suffers from a scale drift problem [2]. Stereo Vision avoids the scale estimation issue but introduces a computational overhead of depth estimation and image synchronization [3]. Microsoft Kinect provides the color-depth (RGB-D) data [4] having benefits of laser and vision sensing together. Although it has shown lot of promise in various robotic application but the limiting factor is it's detection range(3-4 meters) [4]. and it is confined to restricted environement (feature-enriched or structural environments). The research work carried out in 6DOF challenging environments [5, 6] has shown that inertial and vision modalities are attractive solution because of their complementary nature. Inertial sensor are accurate for a short period and helps to bound the pose error in RGB-D approaches (failure cases) whereas visual estimates provide corrections for longer term navigational drift [7].

1.1 Contribution

The motivation behind our work is to make Kinect sensor tackle the challenging structural/environmental problems like insufficient range data (long corridors) or significant loss of visual or geometrical structures. RGB-D approaches usually relies on color plus depth information [4, 8] or depth-only information [9] for pose estimation. Generally pose estimated from RGB-D approaches are robust as they utilize both geometric and visual characteristic [8]. However these approaches are unable to cope with no or insufficient depth data/visual features. The short term vulnerability to these challenging conditions is addressed with the integration of inertial sensor information.



(a) Input Image (b) Sparse 3D map

(c) Point cloud before pose-graph optimization (d) Point cloud after pose-graph optimization

Figure 1: Results of our proposed approach for the lab environment

We presents an IMU aided RGB-D navigation system which is based on both RGB-D and inertial information to obtain consistent metric maps in an on-line probabilistic framework. Fig 1 shows the output of our proposed approach. The complementary fusion of RGB-D and inertial information is done by Extended Kalman Filter (EKF). The prior motion of camera is predicted by IMU and measurement constraints are provided from RGB-D modules. The short term failure cases of RGB-D pose estimator (for example, moving in long corridors or working in non-stationary environment or sudden appearance of planar geometrical structures in the environment) is resolved with inertial prediction. The proposed framework is shown in Fig 2 where front-end estimates the map/ego-motion in real-time whereas back-end runs in off-line to minimize the global inconsistencies of the map (detail discussion in Section 5)

978-1-4673-9073-6/17/$31.00 ©2017 IEEE

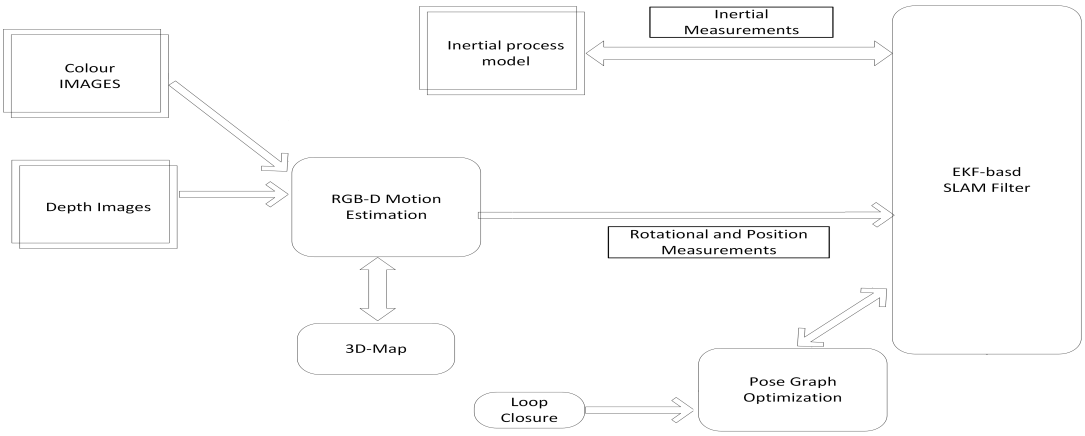


Figure 2: Overall flow chart of proposed SLAM framework

2 Related Work

This section is devoted to RGB-D SLAM and Inertial aided approaches. The recent work on RGB-D SLAM has focused on depth only [9] or depth-color [4, 8] information. The work of [9] utilizes the depth-only information (structural property of the environment). The approaches have limitations, as they are vulnerable to violations of the Kinect failure cases (unstructural environement).

In Visual-Inertial navigation based approaches, the integration methodology can be divided into loosely and tightly coupled approaches. In a tightly coupled system, vision and inertial systems are treated as a single module. However these approaches are computationally expensive (limiting the number of features in the map for realtime performance) [6]. Alternative architecture is loose coupling, which deals the vision and inertial sensor as separate module and implies an EKF at the top hierarchical level receiving only inputs from visual estimator [5]. The advantage of this framework is constant time processing and treatment of visual pose estimator as a black box. Our proposed work is based on loosely couple architecture similiar to [7] work. They perform predictions using the visual pose estimator (stereo camera) and corrections with the inertial sensor. We adopt the opposite approach and predict with the inertial measurements (as the inertial sensor runs at highest frequency and making the overall system to run at higher rates).

Another stream of work in Visual-Inertial aiding deals with aiding the gyro (rotation) information to the Kinect Sensor [10, 11]. In [10] gyro information is used as a initialization for the pose estimator in Iterative closest point (ICP). The short coming of these approaches is lack of full IMU data (accelerometer data for short-term position information for continuous operation of RGB-D system) and reliance on off-the-shelf IMU filtering approach.

Recent work of [1] utilizes the IMU for RGB-D sensor. The motivation of their work is similar to ours that is, operating in environment with fewer features. Their work uses linear kalman filter for translation component and EKF for rotational component for data fusion. Our work differs from them as we formulated the problem as loosely coupled nonlinear data fusion (EKF framework for front-end). We incorporated the global pose graph optimization (back-end) along with EKF filtering for more precised results.

The remainder of the paper is organized as follows: Section 3 briefly explain the probabilistic prediction model based upon inertial sensor. Section 4 will provide an overview of RGB-D based approach. Section 5 provides the detail of EKF update step and pose graph mapping. Section 6 will present the results and discussions followed by Conclusion.

3 EKF based Inertial Prediction

In this work, Inertial sensor is used as a predictor in the EKF whereas feature observations from RGB-D sensor is used in the EKF update. Full 6DOF (degree of freedom) solutions have gained a lot of interest with low-cost inertial sensing [6].

The angular measurements from the gyros (of inertial sensor) are integrated to estimate the attitude. The estimated attitude helps to transform the accelerometer data into the navigation frame (from the body frame). The Inertial Navigation System (INS) is used to calculate the position and velocity from the navigation frame acceleration (after compensating the gravity force) [12]. In the EKF filter the position of the platform is represented as $\mathbf{p}$ whereas the velocity is $\mathbf{v}$ and $\mathbf{q}$ describes the quaternion rotation at discrete time k. Th state vector $\widetilde{\mathbf{x}}$ of EKF filter is defined as:

$$\widetilde{\mathbf{x}}_k \quad = \quad \left[\ \mathbf{p}^n(k), \mathbf{v}^n(k), \mathbf{q}^n(k) \ \right]^T \tag{1}$$

The propagation of the state vector $\mathbf{x}_k$ at time step (k-1) to the next time step (k) is defined by the prediction model as:

$$p\left(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{u_k}\right) \Leftrightarrow \mathbf{x}_k \quad = \quad \mathbf{f}\left(\mathbf{x}_{k-1}, \mathbf{u}_k\right) + \mathbf{w_k} \tag{2}$$

Where the nonlinear process model $\mathbf{f}(.)$ with $\mathbf{w_k}$ as a zero-mean noise of covariance $\mathbf{Q}$. The state vector is evolved as :

$$\widetilde{\mathbf{x}}_k \quad = \tag{3}$$

$$\begin{bmatrix} \mathbf{p}^n(k) \\ \mathbf{v}^n(k) \\ \mathbf{q}^n(k) \end{bmatrix} \quad = \quad \begin{bmatrix} \mathbf{p}^n(k-1) + \mathbf{v}^n(k-1)\triangle t \\ \mathbf{v}^n(k-1) + [(\mathbf{q}^n(k-1) \odot \widetilde{\mathbf{f}}^b(k)) \\ \odot \mathbf{q}^{n*}(k-1) + \mathbf{g}^n]\triangle t \\ \mathbf{q}^n(k-1) \odot \triangle \mathbf{q}^n(k) \end{bmatrix}$$

where the accelerometer measurement is $\bar{\mathbf{f}}^b$, the quaternion conjugate vector is $\mathbf{q}^{n*}(k)$. The input control is $\mathbf{u}$, gravity vector is $\mathbf{g}$ and linear acceleration and angular velocity is $\bar{\mathbf{f}}^b$, $\triangle\mathbf{q}^n(k)$ respectively.

The front hand of the system is based upon the decoupling of the two core modules: the front-end is EKF based state estimation whereas the back-end is pose graph optimization.

4 RGB-D BASED Pose Estimation

Initially from the input 8bit gray scale image the Harris corners [13] are detected. The corners whose respective depth is unavailable are abandon. The pixel position of features u, v and its respective raw depth ρ are converted into 3D feature points as:

$$Z = \frac{z_{ref}}{1 + (\frac{z_{ref}}{fb})\rho} \tag{4}$$

$$X = \frac{Z(u - c_u + \delta u)}{f} \tag{5}$$

$$Y = \frac{Z(v - c_v + \delta v)}{f} \tag{6}$$

Where z_{ref} is the distance of reference plane to the sensor, f is the focal length, (c_u, c_v) are the principal point, b is the baseline in meters and $(\delta u, \delta v)$ are the lens distortion terms.

SURF descriptors [13] are estimated from the detected features for matching purpose. The covariance of each 3D feature is estimated using [14] approach where ρ is treated as random variable with standard deviation as $\sigma\rho$ as:

$$\sigma_Z = \frac{1}{fb} Z^2 \sigma\rho \tag{7}$$

$$\sigma_X = \frac{u}{fb} Z^2 \sigma\rho \tag{8}$$

$$\sigma_Y = \frac{v}{fb} Z^2 \sigma\rho \tag{9}$$

With a 3x3 covariance matrix (Ω) for a 3D feature point is estimated from them.

A map of 3D features with their respective covariance/descriptor is maintained to reduce the effect of drift. A ring buffer is maintained with constant memory limit and behave as a short term memory. The 3D features ($\bar{\mathbf{D}}$) descriptors from Kinect frame are matched with the existing map features ($\bar{\mathbf{S}}$). Sum of absolute difference is used as a matching score. RANSAC is used to estimate the rigid body transformation from the corresponding feature matches using [15] in an iterative fashion. This estimated rigid transformation is provided to weighted ICP as an initial guess for fine refinement as:

$$argmin_{\bar{\mathbf{R}},\bar{\mathbf{T}}} \frac{1}{N} \sum_{i=1}^{N} \Omega^s{}_i ||\bar{\mathbf{S}}_\mathbf{i} - \bar{\mathbf{R}}(\bar{\mathbf{D}}_\mathbf{i}) - \bar{\mathbf{T}}||^2 \tag{10}$$

Where $\{\bar{\mathbf{R}}, \bar{\mathbf{T}}\}$ is the rotation and translation vector obtained from the rigid body transformation.

The feature points ($\bar{\mathbf{D}}$) are updated using the transformation matrix in order to update or add new feature points into the map ($\bar{\mathbf{S}}$). The points are updated if they exist within a defined threshold however other points are included as new points in the map. To fuse the RGB-D information in the EKF filter, the measurement model is defined as:

$$\begin{aligned} p\left(\mathbf{x}_k \mid \mathbf{z}^{Map}\right) \Leftrightarrow \mathbf{z}^{Map} \\ = h^{Map}(\mathbf{x}_k) + \mathbf{r}_\mathbf{k}^{\mathbf{Map}} \\ = \begin{bmatrix} (p^n(t) + \mathbf{C}_{q_b^n} p_c^b) \\ \mathbf{C}_{q_b^n} \end{bmatrix} + \mathbf{r}_\mathbf{k}^{\mathbf{Map}} \end{aligned} \tag{11}$$

Where $\mathbf{C}_{q_b^n}$ is navigation to body transformation obtained from the state vector, p_c^b is the camera position in body frame known and $\mathbf{r}_k^{Map}$ is zero-mean noise with its respected covariance $\mathbf{R}^{Map}$.

For measurement error is defined through innovation vector as:

$$\nu^{\mathbf{Map}} = \left[\widetilde{\mathbf{z}}^{Map} - h^{Map}(\widetilde{\mathbf{x}}_k) \right] \tag{12}$$

where $\widetilde{\mathbf{z}}^{Map}$ is: the estimated rigid body transformation $\{\bar{\mathbf{R}}, \bar{\mathbf{T}}\}$.

$$\widetilde{\mathbf{z}}^{Map} = \left[\bar{\mathbf{T}} \quad \bar{\mathbf{R}} \right]^T \tag{13}$$

5 Pose estimation and Mapping

5.1 EKF Pose estimator

In the featureless or unstructured environment the RGB-D failure cases can happen, hence the inertial sensor helps to predict the state of platform when RGB-D measurements are unavailable, however when the RGB-D pose estimates are available then the RGB-D innovation constraints are used for EKF-update step. The innovation constraints with their respective measurement matrix H (measurement model linearized with respect to state $\mathbf{x}$), we can update the EKF state vector as follows:

- Measurement Innovation Covariance 'S' estimation: S = H P H^T + R
- Gain 'K' estimation: K = P H^T S^{-1}
- State correction step: $\mathbf{x}$ = K $\nu^{\mathbf{Map}}$
- State covariance: P = (I - K H) P $(I - KH)^T$ + K R K^T

The measurement covariance matrix (R) for the RGB-D constraints are estimated from the motion estimation data of ground truth data against the proposed approach. One aspects that is very important during EKF update step is; handling the measurement delay. The occurrence of the delay is usually due to the computational cost of vision processing. We maintain a ring buffer of past states of EKF prediction with time information (a global time-stamp of IMU is used during inertial-vision data acquisition).

5.2 Mapping

Loop closure in SLAM provides a global consistency for locally accurate pose estimator. We adopted the approach of [16] for mapping purpose which is based upon pose graph optimization. The front-end estimates the pose using EKF which are therefore provided to pose graph optimization. A selection of keyframe is based upon the predefined distance travel threshold between image frames . The loop closing is

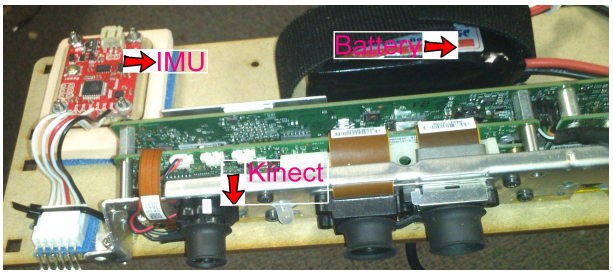


Figure 3: The sensing suit in the experiments containing Batteries, Kinect and Inertial sensor

implemented using the feature point matching (surf descriptors between the current image and the existing keyframes). Each detected loop provide an additional constraint in the pose graph. The optimization is carried out on the pose graph until convergence and the EKF state vector is updated with the optimized pose vector (as a updated measurement).

6 Results

Indoor/outdoor collected datasets are used in the experiments for evaluation purpose. The acquisition of the data is carried out using a sensing suit containing Kinect, Inertial sensor and batteries as shown in Fig 3. The RGB-D frames at QQ-VGA resolution are acquired at 22Hz and Inertial data is acquired at 38Hz. The VICON system is used as a ground truth for indoor experiments at 100Hz (the position accuracy is less than 1cm whereas attitude accuracy is less than $1\circ$). The data acquisition is carried on i3 laptop with Ubuntu as operating system. Robot operating system (ROS) based software synchronization is used for Kinect and inertial sensor.

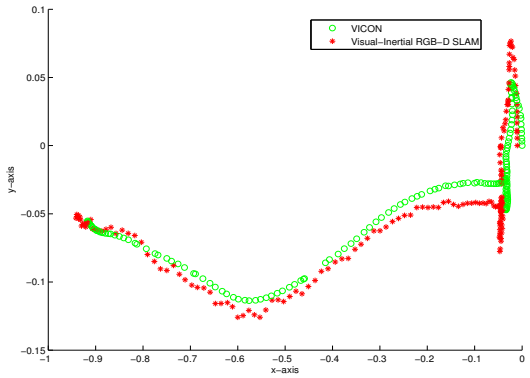
6.1 VICON Evaluation Dataset

The sensing suit is used to acquire data in an indoor environment of $8m \times 6m \times 3m$ with nine VICON cameras (walls are texture-less). On the sensing suit the reflective markers are placed (for the VICON system) to precisely estimate the pose of the platform. The indoor dataset constitutes of 315 RGB-D frames and 520 inertial measurements. Figure 4 shows the positional accuracy of proposed SLAM approach against the VICON system. Fig 5 shows the attitude accuracy of proposed work against the VICON system (the quaternions are converted to Euler angles for visualization purposes).

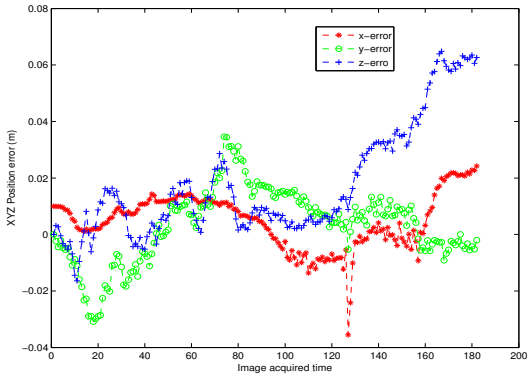
In addition the back-end mapping results can be seen in Fig 1 for office like environment with 200 Kinect frames and 332 inertial measurements.

6.2 Large-scale Dataset

A dataset for an area of $25m \times 12m$ is acquired with 3072 RGB-D frames in a rectangular corridor loop (the corridor dataset contain the failure cases such as planar long corridor with texture-less walls). The RGB-D approach without Inertial aiding shows the drift in camera attitude (i.e. roll, pitch, yaw divergence) whereas the proposed approach has



(a) 2D positional comparison



(b) Positional error x-y-z directions

Figure 4: Positional accuracy of proposed SLAM approach as compare to centimeter level accurate system

bounded the pose drift using gravity vectors and gyro information from Inertial sensor as shown in Fig 6.

7 Conclusion and Future work

This paper presents an inertial aided RGB-D SLAM framework. We have proposed a loosely coupled approach of combining Inertial information with RGB-D sensor. The system consists of real-time front-end and off-line back-end for ego-motion and map estimation. Extensive experimentation is carried out in the indoor environment and evaluated against a very accurate ground truth (VICON) with promising results. The future work is to evaluate the proposed approach on the outdoor large scale dataset.

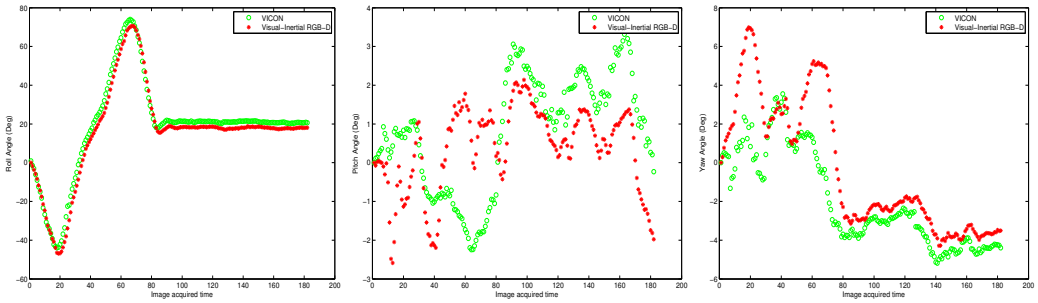
References

[1] N. Brunetto, S. Salti, N. Fioraio, T. Cavallari, and L. Stefano. Fusion of inertial and visual measurements for rgb-d slam on mobile devices. *IEEE International Conference on Computer Vision Workshop (ICCVW))*, 2015.

[2] H. Strasdat, J.M.M. Montiel, and A.J. Davison. Visual slam: Why filter? *Image and Vision Computing (IMAVIS)*, 2012.

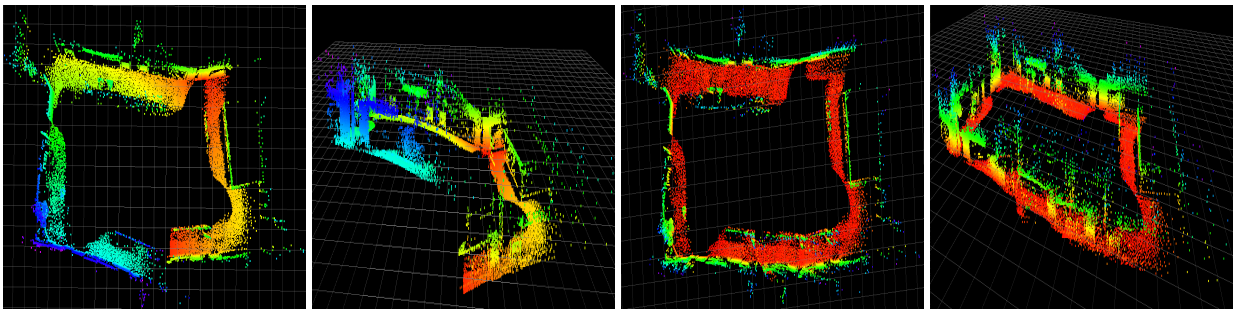
[3] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *Journal of Robotics Research (IJRR)*, 2012.

[4] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. *In Proceedings of Intelligent Robot Systems (IROS)*, 2012.



(a) Roll (b) Pitch (c) Yaw

Figure 5: Attitude accuracy of proposed approach against the VICON system



(a) 2D view RGB-D without IMU aiding (b) 3D view RGB-D without IMU aiding (c) 2D view IMU aided RGB-D (d) 3D view IMU aided RGB-D

Figure 6: Pose drift comparison of RGB-D with/without Inertial sensor

[5] S. Weiss and R. Siegwart. Real-time metric state estimation for modular vision-inertial systems. *In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011.

[6] U. Qayyum and J. Kim. Seamless aiding of inertial-slam using visual directional constraints from a monocular vision. *In Proceedings of Intelligent Robot Systems (IROS)*, 2012.

[7] K. Konolige, M. Agrawal, and J. Sola. Large scale visual odometry for rough terrain. *in In Proc. International Symposium on Robotics Research*, 2007.

[8] J. Stuckler and S. Behnke. Integrating depth and color cues for dense multi-resolution scene mapping using rgb-d cameras. *In Proceedings of Multisensor Fusion and Information Integration (MFI)*, 2012.

[9] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. *In Proceedings of the 24th annual ACM symposium on User interface software and technology, ACM*, 2011.

[10] B. Bouvrie. Improving rgbd indoor mapping with imu data. *Delft University of Technology Masters Thesis in Embedded Systems*, 2011.

[11] H. Ovren, P. Forssen, and D. Tornqvist. Why would i want a gyroscope on my rgb-d sensor?. *Proceedings of IEEE Winter Vision Meetings, Workshop on Robot Vision (WoRV13)*, 2013.

[12] J. Kim and S. Sukkarieh. Real-time implementation of airborne inertial-slam. *Robotics and Autonomous Systems*, 2007.

[13] A. Ess H. Bay, T. Tuytelaars, and L.V. Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 2008.

[14] K. Khoshelham and S O. Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors (doi:10.3390/s120201437)*, 2012.

[15] B. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, 1987.

[16] R. Kummerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. g2o: A general framework for graph optimization. *In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011.