Proceedings of Australasian Conference on Robotics and Automation, 2-4 Dec 2013, University of New South Wales, Sydney Australia

Inertial-Kinect Fusion for Outdoor 3D Navigation

Usman Qayyum and Jonghyuk Kim
Research School of Engineering, Australian National University
{usman.qayyum, jonghyuk.kim}@anu.edu.au

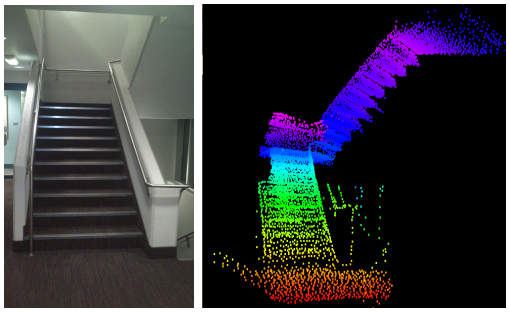
Abstract

The lightweight and low-cost 3D sensing device, such as Microsoft Kinect, has gained much attention in computer vision and robotics community. Although quite promising and successful for indoor applications, its outdoor usage has been significantly hampered by its short detection range (around 4 meters) coupled with ambient infrared interference. This paper addresses the theoretical and practical development of an Inertial-Kinect fused SLAM framework that can handle the 3D to 2D degeneration in Kinect sensing, called a depth dropout problem. The vision node is designed to provide either full 6DOF or partial 5DOF vehicle pose measurements depending on the depth availability, whilst the low-cost inertial system designed in house (less than $40AUD) enables continuous metric mapping and navigation. Indoor and outdoor experiment results are provided, demonstrating the robustness of the proposed approach in a challenging environmental conditions.

1 Introduction

There has been significant progress in achieving self-contained navigation under challenging environments in the last couples of years, spanning from the radio-signal aided navigation to on-the-fly map integrated systems such as simultaneous localization and mapping (SLAM). The map generation has been done by incorporating various perception sensors such as laser or vision. The visual sensing modality in particular has drawn great attention due to its passiveness, lightweight and rich information content, coupled with the advances in embedded computing technology.

A monocular camera has been extensively studied for navigation and mapping purpose by many researchers. The projective nature and lack of scale information of the environment make it very challenging to avoid scale drift [Strasdat et al., 2012] or fuse it with other metric sensors. Stereo vision avoids the arbitrary scale assumption or drift problem but introduces a computational overload of depth calculation as well as image synchronization.



(a) Input image from staircase sequence (b) Generated map

Figure 1: 3D map generated from our proposed approach while walking through the staircases

The recent success of low-cost 3D sensing devices, such as the Microsoft Kinect, has drawn much attention in the research community [Henry et al., 2012]. The acquired colour and depth (or RGB-D) data provides the many benefits of both laser (depth) and vision (color image) modalities in a single package. The real-time accessibility of three dimensional point clouds from the Kinect has been used for navigation, path planning and collision avoidance [Henry et al., 2012; Endres et al., 2012].

Although quite promising, there are still many challenging environmental problems for Kinect based SLAM:

- *Partial or no depth information*: Depth data from the Kinect sensor is unavailable or partially available for challenging environments (such as long corridors, big open spaces or outdoors) due to limited range and infra-red sunlight interference. The presence of partial depth data degrades the quality of

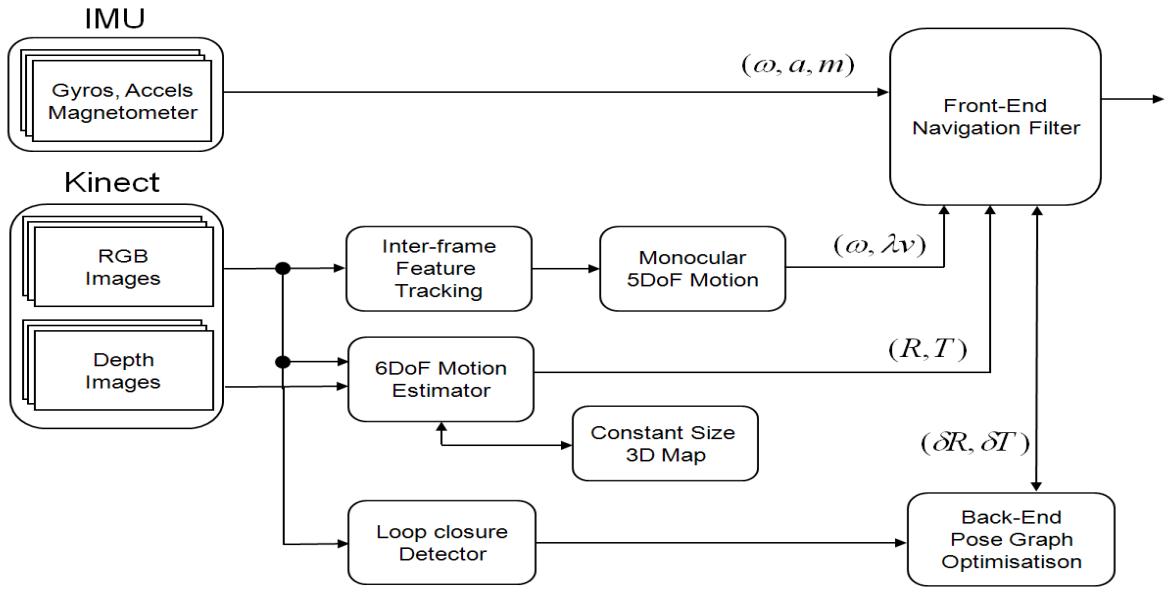


Figure 2: Flow chart of our approach: EKF based front-end and Pose-graph based back-end system

the estimated pose and unavailability of depth data makes the existing RGB-D approaches inapplicable for outdoor environments.

The problem of insufficient or no depth data from Kinect sensor is called here *a depth dropout problem*, hence making RGB-D sensor act as monocular camera. This work addresses the depth dropout problem by proposing a novel inertial-Kinect fusion framework that can handle both 3D and 2D sensing modes. The contributions of this work are:

- The vision node is designed in a modular way that can provide either full 6DOF (rotation and translation) or partial 5DOF (rotation and scale-ambiguous translation) vehicle pose measurements during depth dropouts.
- The visual translation in 5DOF mode is treated as a *directional constraint* of the motion rather than estimating the depth from inertial output. Consequently the vision output can be seamlessly fused with the low-cost inertial system without delay or until the depth to converge. The depth can be resolved, whilst aiding inertial sensors, until the camera has enough parallax motion.

Figure 1 shows the point-clouds outputs from the proposed approach. The dynamic motion of camera is predicted by inertial sensor and updated using the RGB-D or monocular measurements. Figure 2 illustrates the overall system architecture where the vision node provides either RGB-D or monocular information, depending on the availability of the depth information. The proposed framework and experimental results makes the RGB-D sensor a viable solution for indoor and outdoor navigation, which has never been demonstrated before, to the best of authors knowledge.

The outline of this paper is as follows: Section 2 will provide related work in RGB-D sensing and navigation. Section 3 will discuss the inertial system node and Section 4 will detailed a RGB-D sensing node in a modular architecture and a measurement model for monocular odometry constraint. Section 5 will provide the pose estimator and mapping details. Results and discussions will be presented in Section 6 followed by Conclusion.

2 Related Work

Previous work on RGB-D based SLAM has used full depth-colour as in [Endres et al., 2012] or depth-only information [Izadi et al., 2011]. In the depth-colour case, image features are detected and their corresponding depths are used to build 3D feature points. The feature points are then matched with the key-frame features to estimate the camera pose. The work is based upon a multi-threaded approach and pose graph optimization is performed for global consistency of the map as an off-line process. The work of [Izadi et al., 2011] relies on depth-only information and hence used the Iterative Closest Point (ICP) approach to exploit the structural properties of environment. Another research work has been dedicated to using depth information in the monocular SLAM framework [Scherer et al., 2012]. The depth information is used to solve the drawbacks of the monocular parallel tracking and mapping (PTAM) algorithm, for instance automatic bootstrapping and 3D feature ini-

tialization. Recently there is another work by [Whelan et al., 2013] in which 3D visual odometry is integrated with the ICP-based SLAM approach.

The aforementioned state-of-the-art techniques rely heavily on the availability of the depth data and hence are deemed to fail at depth dropout cases. The depth dropout issue is recently considered by [Gibson et al., 2012], in which they treated this problem as an off-line optimization problem in an indoor environment. The approach is based upon heuristically combining monocular SLAM and RGB-D SLAM into an off-line local mapping problem. In this work we focus on the on-line integration approach of combining monocular or RGB-D within inertial SLAM framework aiming for outdoor navigation.

3 IMU-based Pose Estimation

An inertial measurement unit (IMU) is comprised of triaxial gyros/accelerometers arranged orthogonally and providing 6DOF measurements of the platform attached. The procedure used to estimate the position and velocity of the platform in the navigation frame is called the INS algorithm [Kim and Sukkarieh, 2007]. The rotational rate measured from the gyros in the body frame is integrated to form the attitude (or orientation) of the vehicle. The attitude is then used to transform the accelerometer measurement into the navigation frame quantity, which is then double integrated to get the position.

Let the motion of the vehicle be described by the state equation $x = f(x, u)$ with the state vector x consisting of position (or translation) T, velocity V and rotation R in the navigation frame, and the control input u with the acceleration a^b and the angular velocity ω^b in the body frame.

Note that $T, V \in \mathbb{R}^3$ and $R \in SO(3)$. The state equations can then be written as

$$\begin{aligned} \dot{T} &= V \\ \dot{V} &= R\, a^b + g \ , \\ \dot{R} &= R\, [\omega^b]_\times \end{aligned} \tag{1}$$

where $g = [0, 0, -9.8m/s^2]^T$ is the gravity vector and $[\omega^b]_\times$ is a skew-symmetric form of the angular velocity vector.

4 Visual Pose Measurement

The pipeline for the RGB-D pose estimator is to estimate 1) the full 6DOF motion estimation when the depth data is available, or 2) the partial 5DOF motion when the depth dropouts occur as illustrated in Figure 2.

4.1 6DOF Pose Measurement

Harris corners [C. Harris and M. Stephens , 1988] are initially detected on the gray scale image from the Kinect frame. The corner features for which the corresponding depth is unavailable are discarded. The spatial location of the feature in the pixel coordinates with depth gives $(u, v, d) \in \mathbb{R}^3$, which can be converted into 3D Euclidian feature position, $(x, y, z) \in \mathbb{R}^3$, relative to the camera. The mapping function $g : (u, v, d) \rightarrow (x, y, z)$ becomes:

$$\begin{aligned} x &= \frac{z}{f}(u - u_0) \\ y &= \frac{z}{f}(v - u_0)) \\ z &= \frac{f}{d}\, b, \end{aligned} \tag{2}$$

where f is the camera focal length, (u_0, v_0) being the centre of the image, and b the baseline between the infrared emitter and the infrared camera. The intrinsic calibration of Kinect is assumed to be known. The related covariance matrix of the transformed Euclidian 3D position $\mathcal{W}$ can be computed by using Jacobians of the mapping and assuming independent noises in pixel and depth measurements

$$\begin{aligned} \mathcal{W} &= J \begin{pmatrix} \sigma_u^2 & 0 & 0 \\ 0 & \sigma_v^2 & 0 \\ 0 & 0 & \sigma_d^2 \end{pmatrix} J^T \\ \text{where, } J &= \frac{\partial g(x, y, z)}{\partial (u, v, d)}. \end{aligned} \tag{3}$$

Speed-up-robust-features (SURF) descriptors are used for the purpose of feature matching [Bay et al., 2008]. In order to reduce the short term drift, we maintained a map of 3D features with their covariance and descriptors in a ring buffer. The ring buffer has a constant memory size and thus enables the front-end to perform in constant time.

The first Kinect frame features are declared as part of the map ($\mathbf{M}$) defined in the local navigational frame. All the subsequent feature measurement data ($\mathbf{D}$) are matched with the existing map features using the SURF descriptors. The comparing score is based on the sum-of-absolute-difference and if the score is within a set threshold then it is declared as a matched-pair. As this matching can still lead to wrong matches, RANSAC is used to remove the outliers.

The rigid body transformation (R, T) of the vehicle is obtained in two steps. First an initial estimate is obtained using the closed-form solution as in [Besl and Mckay, 1992]. It is then used to run a weighted ICP for fine refinement.

$$\arg\min_{R,T} \left(\frac{1}{n} \sum_{i \in \mathcal{A}} \mathcal{W}_i |\mathbf{M}_i - (R\, \mathbf{D}_i + T)\,|^2 \right), \tag{4}$$

where i is the index of inlier feature set $\mathcal{A}$ and $\mathcal{W}$ is the weighting matrix given in (3).

The rigid body transformation is used to transform the feature data with their associated covariance to update or add new features into the map. The points that are within a predefined Euclidean vicinity are declared as update-points, where others are declared as new-points. The existing points are updated using a weighted averaging method whilst the inverse covariances are added together. The feature descriptors present in the map are updated with the new ones (a different update scheme can be employed and researched but that was not the focus of this work). The new-points are added to the map with their descriptors and covariances. If the limit of ring buffer is reached then the old features are deleted. Retaining the most spatially informative features in the map instead of deleting them, can further enhance the performance (which is currently under investigation).

4.2 5DOF Pose Measurement

When the depth information is not available due to either the features are beyond the detection range or intensive sunlight interference, the RGB-D sensor degenerates into a monocular camera. In this case, the vision system can still provide 5DOF pose information: the rotation and translation with an unknown scale $(R, \lambda T)$. There have been three approaches to deal with this unknown scale:

- The pose can directly be used for navigation but in the unknown scale-space as in many pure vision-based approaches such as Visual-SLAM or RatSLAM [Milford and Wyeth, 2008].
- The unknown scale value can be recovered by observing any a prior infrastructural information such as visual markers or known vehicle heights [Kim and Sukkarieh, 2007].
- The unknown scale can be estimated using other aiding sensors such as wheel or inertial sensors, which integrate the wheel speed or acceleration to provide the scale information [Weiss and Siegwart, 2011].

The third option seems most relevant to our system. However it was observed that the double integrated acceleration from the low-cost inertial sensor can diverge quite quickly before the camera experiences enough parallax for the depth resolution.

In this work we treat the visual translation with the unknown scale as *a directional constraint of the motion*. This enables the direct fusion of inertial and vision output without relying on the scale estimation [Qayyum and Kim, 2012].

The feature matching and detection phase of the monocular pipeline is similar to RGB-D pipeline except that the depth information is not available. In this pipeline we do not maintain the feature map, and a frame-to-frame visual odometry is extracted.

Harris features and SURF descriptors are extracted from the gray scale images. The feature points from the current image are matched with those of the previous image using a 5-point algorithm with RANSAC processing [Hartley and Li, 2012]. The inliers are then used to extract delta rotation (ΔR) and delta translation $(\lambda \Delta T)$ over two consecutive images. Therefore, by using the sampling time, the vision node outputs rotational rate and translational velocity (up to scale) $(\omega, \lambda v)$ in the camera coordinates. Without loss of generality, we assume the camera and body coordinates are aligned each other. To fuse these with the inertial sensor outputs, which operate in a metric space, both inertial and visual velocities are converted into unit velocities. For instance, the inertial unit velocity $[\hat{v}_x, \hat{y}_x, \hat{v}_z]^T$ in the body coordinates becomes

$$\begin{bmatrix} \hat{v}_x \\ \hat{v}_y \\ \hat{v}_z \end{bmatrix}^b = \frac{1}{\sqrt{V_x^2 + V_y^2 + V_z^2}} \begin{bmatrix} V_x \\ V_y \\ V_z \end{bmatrix}^b_{IMU}. \tag{5}$$

If the vehicle motion is constrained to the ground, the so called nonholonomic constraint could be applied [Dissanayake et al., 2001]. For example the unit y and z velocity components are set to zero $(\hat{v}_y = 0, \hat{v}_z = 0)$. In the general case, such as a hand-held camera, this type of constraints cannot be applied. However the vision system can still provide such motion constraints: the lateral and normal velocities are constrained by the visual velocity

$$\begin{bmatrix} \hat{v}_x \\ \hat{v}_y \\ \hat{v}_z \end{bmatrix}^b_{IMU} = \begin{bmatrix} \hat{v}_x \\ \hat{v}_y \\ \hat{v}_z \end{bmatrix}^b_{Vision} + \begin{bmatrix} \nu_x \\ \nu_y \\ \nu_z \end{bmatrix}. \tag{6}$$

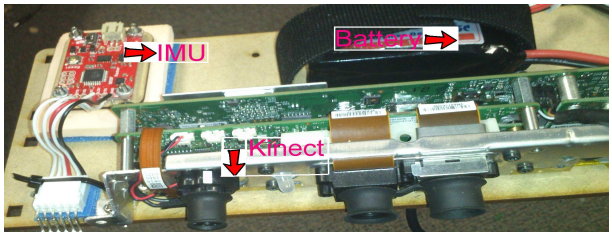
where $[\nu_x, \nu_y, \nu_z]^T$ is a measurement noise vector.

5 Integrated Navigation and Mapping

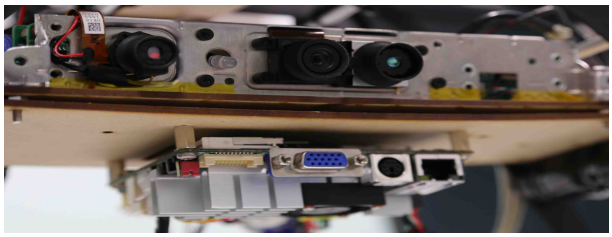
The integrated fusion system consists of 1) a real-time front-end system based on extended Kalman filtering framework and 2) a pose-graph based optimisation for the global consistency. We followed a loosely coupled architecture as shown in Figure 2 in which features are not maintained in the main integration filter [Weiss and Siegwart, 2011]. Although this is a suboptimal approach when compared to the tightly coupled system, we adopted this to avoid the well known scalability issue in SLAM [Qayyum and Kim, 2012]. Another advantage is the vision based pose estimator becomes modular and can be treated as a black box system, thus easily incorporating other vision algorithms. The front-end system is based upon two decoupled modules: the main fusion filter and Kinect-based pose estimator. The back-end side

Table 1: **Evaluation of proposed approach against VICON motion estimates (RMSE error)**

Method	T_x (m)	T_y (m)	T_z (m)	**Roll** (deg)	**Pitch** (deg)	**Yaw** (deg)
Inertial Kinect without monocular constraints	3.25	1.55	3.24	2.124°	3.917°	5.641°
Inertial Kinect with monocular constraints	0.018	0.06	0.13	0.388°	0.647°	0.802°



(a) Top View



(b) Bottom View

Figure 3: Payload system containing power source, Kinect,IMU and Dual-core Atom board

is based upon the pose graph optimization to maintain the global consistency of the system.

5.1 Real-time Front-end Processing

Given that the Kinect dropout can happen, we allow our filter update step to dynamically switch between the monocular or RGB-D measurements. The switching criteria is based upon the presence of the depth features and their spatial geometry. If the number of depth features are greater than the set threshold and the distribution of the features is uniformly spread, then RGB-D measurements are used to update the filter. Otherwise monocular measurements are utilised. The measurement uncertainty related to the visual pose output is dynamically scaled with the number of inliers to gauge the quality of the RGB-D/monocular motion estimation. The measurement delay involved in the vision processing should also be handled carefully within the integration filter. The measurement update has to be synchronized in time with the inertial based prediction step. Therefore we maintain time stamps with the predicted states within the ring buffer. When the measurement is obtained, the past EKF state with matching time is retrieved and updated accordingly. The corrected state is then propagated to the current state.

5.2 Back-end Processing

Loop closure within the vision node provides a global consistency of the system. A common approach is to represent the pose as a node and their relations as edges in a pose graph. Pose graph optimization can be faster than the bundle adjustments which optimizes both states and features but comparatively less consistent [Kummerle et al., 2011].

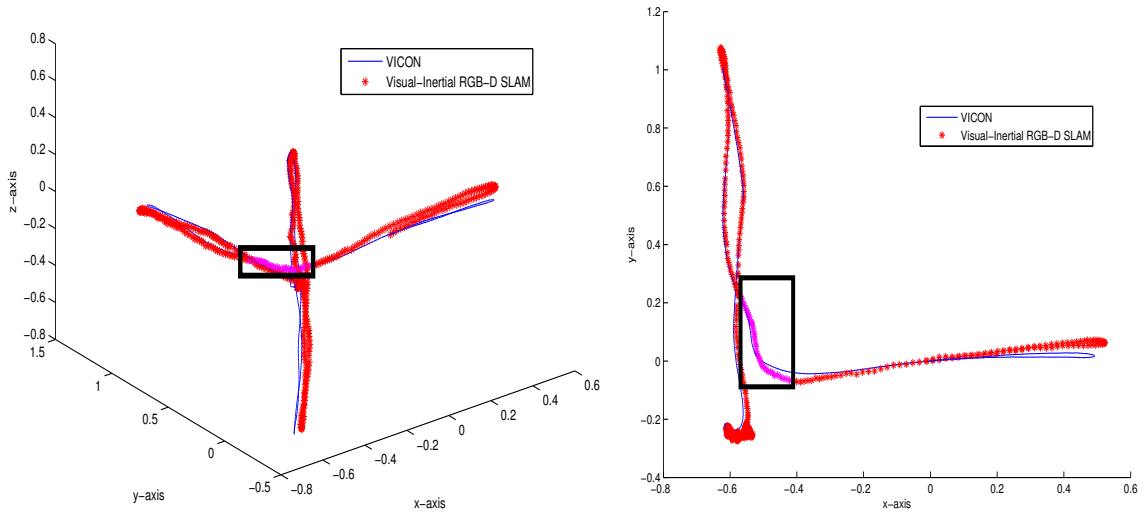
In this work, we have adopted the keyframe-based pose graph optimization method as in [Kummerle et al., 2011]. The pose estimated from the EKF is fed to the back-end graph optimizater. Keyframe selection mechanism is carried out by thresholding the accumulated current motion. Loop closure is detected using the matching of the SURF feature descriptors (detected earlier in the front-end side) between the current image and the existing keyframes. Once the loop is detected a new constraint (rigid transformation) is added to the pose graph and subsequently optimized till convergence. On convergence the EKF state (against the measured time-stamp) is updated with the help of the state ring buffer.

6 Results

The proposed approach is evaluated using an indoor and outdoor dataset. The dataset for the proposed algorithm was collected using a self-powered Inertial-Kinect payload system shown in Fig 3. The dataset includes the Kinect frames (RGB-D) taken at 22Hz and Inertial data acquired from the low-cost IMU at 38Hz. The ground truth data (VICON) at 100Hz (of less than 1cm/1° error) was also acquired for evaluation purposes in the indoor environment. All the data was acquired on a dual-core Atom board running ubuntu with Robot Operating System (ROS) framework. All the data was synchronized with the time-stamp generated by the computer for the respective sensors whereas a ring-buffer approach was used to cater the difference in acquisition time and processing delays.

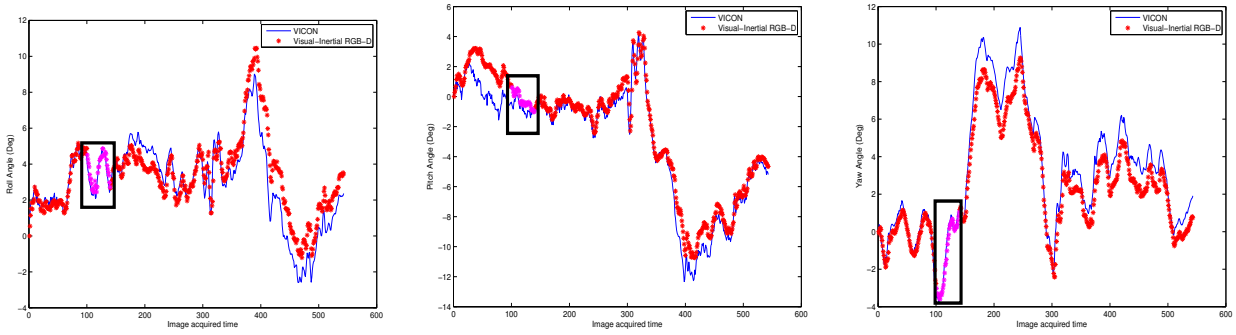
6.1 Indoor Dataset

To fully evaluate the proposed idea, a comparison of the motion estimates of the proposed approach is provided against the VICON ground truth. The payload system was moved in a $8m \times 6m \times 3m$ indoor environment with nine VICON cameras mounted on the three corner of the ceiling walls. We placed five reflective markers asymmetrically and rigidly onto the payload system for motion



(a) 3D view for XYZ axis (in meters) (b) XY view (in meters)

Figure 4: Trajectory comparison of proposed approach against VICON system (the rectangular box shows the use of monocular constraints at depth dropout case)



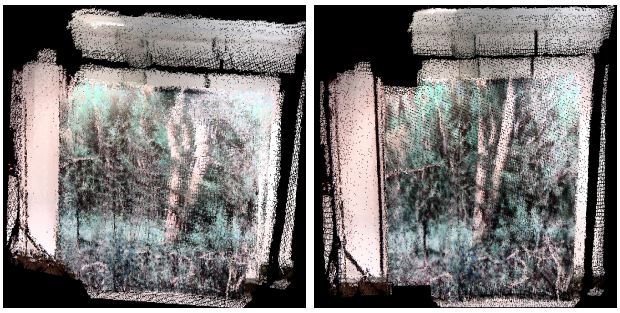
(a) (b) (c)

Figure 5: Attitude comparison of proposed approach against VICON system (the rectangular box shows the use of monocular constraints at depth dropout case)

capturing. The dataset consists of 552 Kinect frames and 800 IMU packets for an indoor environment where one of the wall was textured with forest like environment. To simulate the depth dropout case, we intensionally drop the depth data for some consecutive Kinect frames to evaluate the robustness of our proposed approach.

Figure 4 shows the translational comparison of proposed approach against the ground truth data whereas the depth dropout cases are highlighted via a rectangular boxes (during which monocular directional constraints are utilized by the proposed approach for motion estimation). Fig 5 shows the attitude comparison of the proposed approach with the ground truth where Euler angles are used for visualization purposes.

We have also compared the motion estimates of the proposed approach against the Inertial-Kinect case (without monocular constraints) and used Root Mean Square Error(RMSE) as a quality metric. The RMSE was calculated for the estimated motion estimates from the ground truth data as shown in Table 1. It can be observed that the motion estimates from the proposed approach closely matches with the ground truth data (with RMSE of less than 1 meters/deg). In addition to evaluating the front-end motion estimates, the back-end mapping was applied on the selected key-frames (detailed in section 5.2), to further minimize the drift as shown in



(a) Before optimization (b) After optimization

Figure 6: Before and After pose graph optimization for the indoor VICON dataset where the room wall was textured with forest like environment

Fig 6 (before and after pose graph optimization).

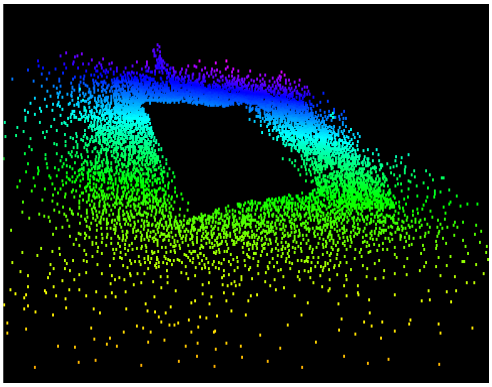
6.2 Outdoor Dataset

The outdoor data was collected from a dense forest where depth dropouts happen due to limited range and ambient infra-red interference. The environment was challenging for SLAM algorithms due to lack of GPS aid and unstructured nature as shown in Fig 7. The dense nature of the forest environment provides a good testbed for our framework to be evaluated against the RGB-D and depth dropout cases (due to covered).

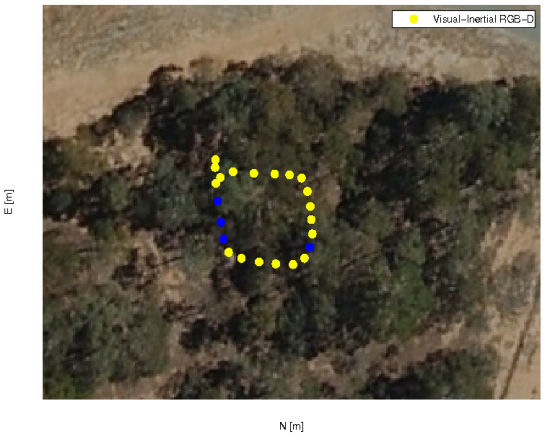
An area of $20m \times 20m$ was explored with 2072 RGB-D Kinect frames in a rectangular loop. 700 depth frames are partially or completely unavailable due to range-limitation or sunlight interference. The estimated trajectory was mapped onto the Google imagery as shown in Fig 8 (in which 600 and 100 are two consecutive patches of depth dropout). The depth dropout cases are handled with the 5DOF monocular constraints and it complements the Inertial-Kinect information to obtain metric maps.

The results shown here demonstrate that the proposed algorithm is capable of operating in different challenging environment, and producing very accurate motion estimates against collected VICON data.

Figure 7: Highly unstructured forest environment



(a) Generated map



(b) Estimated trajectory (the blue-color estimated trajectory points show the use of monocular constraints)

Figure 8: Estimated trajectory and map of the proposed approach

7 Conclusion and Future Work

State of the art approaches in RGB-D SLAM rely heavily on the availability of depth data and hence deem to fail at depth dropout cases. This paper addresses the theoretical and experimental development of an Inertial-Kinect framework to handle the challenging environments. We have proposed an on-line probabilistic approach by combining inertial system with full 6DOF pose or partial 5DOF from the Kinect sensor. The front-end of the system was designed in a modular fashion by decoupling the state estimator and Kinect-based pose estimator (RGB-D and Monocular). The back-end of the system was based upon the pose graph optimization. Indoor and outdoor results demonstrated the the robustness of the proposed framework.

The future work includes applying the navigation outputs for the aerial vehicle control, such as station keeping or trajectory following, under cluttered forest environment.

Acknowledgments

This work is supported by the ARC DP Project (DP0987829) funded by the Australian Research Council (ARC).

References

[Strasdat et al., 2012] Strasdat H. , Montiel J.M.M. and Davison A.J. *Visual SLAM: Why Filter?* . Image and Vision Computing (IMAVIS) 2012.

[Henry et al., 2012] Henry P., Krainin M., Herbst E., Ren X. and Fox D. *RGB-D Mapping: Using Kinect-style Depth Cameras for Dense 3D Modeling of Indoor Environments.* Journal of Robotics Research (IJRR), 2012.

[Endres et al., 2012] Endres F., Hess J., Engelhard N. , Sturm J., Cremers D., and Burgard W. *An evaluation of the RGB-D SLAM system* . In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), May 2012.

[Izadi et al., 2011] Izadi S., Kim D., Hilliges O., Molyneaux D., Newcombe R., Kohli P., 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 ACM symposium on User interface software and technology, ACM, 2011.

[Scherer et al., 2012] Scherer, S.A., Dube, D. and Zell, A., *Using depth in visual simultaneous localisation and mapping.* In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 5216 - 5221, 2012.

[Whelan et al., 2013] Whelan T., McDonald J., Johannsson H., Kaess M. and Leonard J., *Robust Real-Time Visual Odometry for Dense RGB-D Mapping.* In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2013.

[Gibson et al., 2012] Gibson H., Shoudong H., Liang Z., Alen A. and Gamini D., *A robust RGB-D SLAM algorithm* . In Proceedings of Intelligent Robot Systems (IROS), pp. 1714-1719, 2012.

[Qayyum and Kim, 2012] Qayyum, U. and Kim, J., *Seamless aiding of inertial-slam using Visual Directional Constraints from a monocular vision* . In Proceedings of Intelligent Robot Systems (IROS), pp. 4205-4210, 2012.

[Kummerle et al., 2011] Kummerle R., Grisetti G., Strasdat H., Konolige K., and Burgard W., *g2o: A General Framework for Graph Optimization..* In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2011.

[Dissanayake et al., 2001] Dissanayake G., Sukkarieh S., Nebot E., Durrant-Whyte H., *The aiding of a low-cost strapdown inertial measurement unit using vehicle model constraints for land vehicle applications.* IEEE Transaction on Robotics, 17(5):731-747, 2001

[Kim and Sukkarieh, 2007] Kim J., and Sukkarieh S., *Real-time Implementation of Airborne Inertial-SLA.* Robotics and Autonomous Systems, vol. 55, pp. 62-71., 2007.

[Hartley and Li, 2012] Hartley R. and Li H., *Five-point algorithm made easy.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2012.

[Besl and Mckay, 1992] Besl P.J. and McKay N.D., *A method for registration of 3D shapes.* IEEE Transactions Pattern Analysis and Machine Intelligence,14:239256, 1992.

[C. Harris and M. Stephens , 1988] Harris C. and Stephens M., *A combined corner and edge detector.* Proceedings of the 4th Alvey Vision Conference. pp. 147151, 1988.

[Bay et al., 2008] Bay H., Ess A., Tuytelaars T. and Gool L.V., *SURF: Speeded Up Robust Features.* Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346–359, 2008.

[Weiss and Siegwart, 2011] Weiss S. and Siegwart R., *Real-Time Metric State Estimation for Modular Vision-Inertial Systems.* In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2011.

[Milford and Wyeth, 2008] Milford M., Wyeth G., *Single Camera Vision-Only SLAM on a Suburban Road Network.* In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2008.