A FREQUENCY-DRIFT COMPENSATED CLOSED-FORM SOLUTION FOR STEREO RGB-D MAPPING

S.J. Tang [1,3]*, Q. Zhu[2], W. Chen [3], W.X. Wang [1], Y. Li[1], W. Darwish [3], W.B. Li [3]

[1]Research Institute for Smart Cities & Shenzhen Key Laboratory of Spatial Information Smart Sensing and Services, School of Architecture and Urban Planning, Shenzhen University, Shenzhen, PR China - shengjuntang@szu.edu.cn; wangwx@szu.edu.cn; liyou@szu.edu.cn
[2] Faculty of Geosciences and Environmental Engineering of Southwest Jiaotong University, Chengdu, China -zhuq66@263.net
[3] Department of Land Surveying & Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China - wu.chen@polyu.edu.hk; w.darwish@connect.polyu.hk; liwbiz@gmail.com

KEY WORDS: Stereo camera, Camera tracking, Point Cloud, Indoor Mapping

ABSTRACT:

In this work, we present a frequency-drift compensated (Fd-C) closed-form solution for stereo RGB-D SLAM. The intrinsic parameters for each sensor are first obtained with a standard camera calibration process and the extrinsic orientation parameters achieved through a coarse-to-fine scheme that solves the initial exterior orientation parameters (EoPs) from control markers and further refines the initial value by an iterative closest point (ICP) variant minimizing the distance between the RGB-D point clouds and the referenced laser point clouds. With the assumption of fix transformation between the frames with the same timestamp, we define one sensor as reference sensor and the other sensor as slave sensor and the slave frames can be mapped to the timeline of the references sensor. Rather than endow the camera pose of the nearest frame to the slave frames, we derive the accurate camera pose for slave frames in a spatially variant way. Therefore, the pose relations between the slave frame and the adjacent reference frame can be derived, which provided opportunity to use the more accuracy observations from multiple frames for better tracking and global optimization. We present the mathematical analysis of the iterative optimizations for pose tracking in multi-RGB-D camera cases. Finally, the experiments in complex indoor scenarios demonstrate the efficiency of the proposed multiple RGB-D slam algorithm.

1. INTRODUCTION

Creating detailed 3D maps of indoor environments is critical for mobile robotics applications, including indoor navigation, localization and path planning. Simultaneous localization and mapping (SLAM) is key to reliable 3D maps, as it estimates the camera pose accurately, regardless of sensors (Thrun, 2002). Recently, the widespread availability of RGB-D sensors (such as Kinect and Structure Sensor devices) has led to rapid progress in indoor mapping. These tools have several advantages. They are low cost, lightweight and highly flexible and are capable of high-quality 3D perception (Endres et al., 2012; Newcombe et al., 2011). The accuracy of the 3D maps produced by RGB-D devices is highly dependent on the accuracy of the frame registration. RGB-D SLAM systems can be categorized into two types based on the registration method: the dense style and the sparse style (Darwish et al., 2019; Kerl et al., 2013; Mur-Artal and Tardos, 2017; Tang et al., 2018; Tang et al., 2016; Whelan et al., 2012). In order to achieve more robust pose tracking and mapping of visual SLAM, the robotics researcher has recently shown a growing interest in utilising multiple camera, which is able to provide more sufficient observations to fulfil the frame registration and map updating tasks. This implies that better pose tracking robustness can be achieved by extending monocular visual SLAM to utilise measurements from multiple cameras. (Yang et al., 2015) proposed a visual SLAM method using multiple RGB-D cameras, which integrate the observations from multi-camera for camera tracking. Their experiments results implied that dual-Kinects SLAM provided better pose tracking performance than the results from single Kinect. However, they ignored the time-drift between the frames obtained by different cameras, which may result at inaccurate positions of observation used for map updating. Besides, loop closure detection was not been implemented. (Chen et al., 2018) constructed a multiple RGB-D system with three Kinects V2 camera. Three Kinects V2 are mounted on a rig with different directions and synchronized by OpenKinect driver. However, this work mainly concentrated on the intrinsic and extrinsic calibration and verify the effectiveness of mapping using multiply RGB-D cameras.

In this paper, we present a frequency-drift compensated closed-form solution for multiply RGB-D SLAM, which is enable to eliminate the influence of time-drift between different camera during motion tracking. The intrinsic parameters for each sensor are obtained with a standard camera calibration process and the extrinsic orientation parameters achieved through a coarse-to-fine scheme that solves the initial exterior orientation parameters (EoPs) from sparse control markers and further refines the initial value by an iterative closest point (ICP) variant minimizing the distance between the RGB-D point clouds and the referenced laser point clouds. While in theory, a fix rigid transformation should be sufficient to register the frames with the same timestamp from two sensors. Since synchronising multiply Kinect sensors is impossible, there existed frequency-drift due to different topic publish rate of sensors. Then with the assumption of fix transformation between the frames with the same timestamp, we define one sensor as reference sensor and the other sensor as slave sensor and the slave frames can be mapped to the

* Corresponding author

timeline of the references sensor. Rather than endow the camera pose of the nearest frame to the slave frames, we derive the accurate camera pose for slave frames in a spatially variant way. For each slave frame, we make a hypothesis that there exist a corresponding reference frame with the same timestamp and two adjacent frames can be found for each fictitious frame. A linear basis is imposed on the translation and rotation to recover the camera pose of the fictitious frame. A scale parameter is computed from the time interval between the fictitious frame and the adjacent frames. While trilinear interpolation is used to interpolate translation quantities, rotations have to be interpolated over the sphere to achieve constant-speed motion. This is achieved by the slerp operation. Therefore, the pose relations between the slave frame and the adjacent reference frame can be derived, which provided opportunity to use the more accuracy observations from multiple frames for better tracking and global optimization. Finally, the experiments in complex indoor scenarios demonstrate the efficiency of the proposed multiple RGB-D slam algorithm.

Figure 1. Left: Stereo RGB-D mapping system setup, the two Kinects mounted on a Jetson Tx2, one facing upwards and the other one facing downwards. Middle: Camera views of two camera at the same time-stamp. Right: The built dense point cloud using the two Kinect

2. CAMERA CALIBRATION PROCEDURE

Since the stereo cameras in our multiple RGB-D system are mounted with few overlapping in their FOVs, we solve the initial EoPs of two cameras by the corresponding points detected from the chessboard. We defined the downward Kinect as the reference camera C^r and the upward Kinect as slave camera C^s. As shown in Figure 2(a), two sets of frame pairs are collected in different position and direction, and the corresponding point pairs in the chessboard are detected automatically, P^r and P^s represent the corner points in reference sensor and slave sensor respectively. Therefore, an initial transformation T^{INI} between downward and upward cameras can be derivate by minimizing the distance between P^r and P^s.

$$P^r = T^{INI} * P^s \tag{1}$$

As the overlapping region is two small to obtain an accurate EoPs. We further refine the results by an ICP variant minimizing the distance between the RGB-D point cloud and the referenced laser point clouds P^L. We first register the laser point clouds to the point clouds from reference sensor. Since the ICP processing requires a fine initial transformation. Therefore, based on the initial transformation of C^r and C^s, we first transform the point cloud from slave sensor to the coordinate system of reference sensor, then the transformation can be refined by an ICP variant minimizing the distance between the transformed point cloud from slave sensor and the laser point clouds.

$$P^L = T^{Refin} * P^{s\prime} \tag{2}$$

Based on Equation (1) and Equation (2), accurate external orientation parameters of reference sensor and slave sensor can be derived as following.

$$C^r = T^{Refin} * T^{INI} * C^s \tag{3}$$

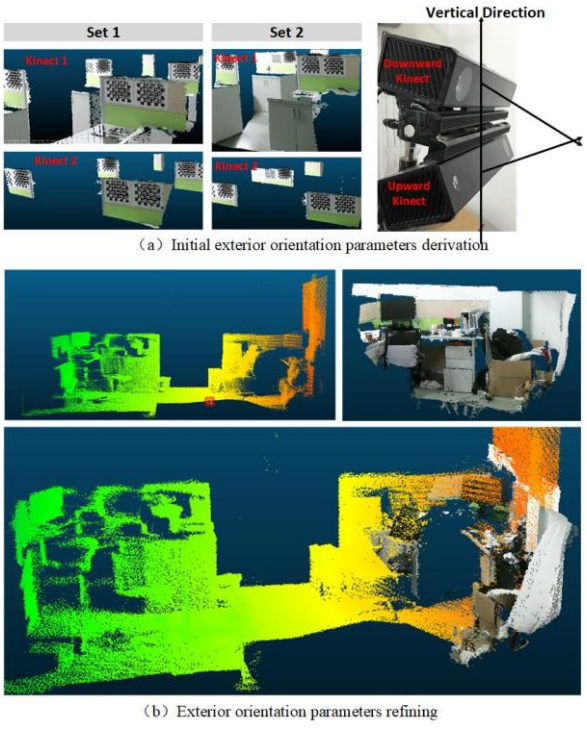

Figure 2. Exterior orientation calibration for stereo cameras

3. FREQUENCY-DRIFT COMPENSATED MULTIPLY CAMERA TRACKING

3.1 Fd-C strategy

Theoretically, a fix rigid transformation should be sufficient to register the frames with the same timestamp from two sensors. Since synchronising multiply Kinect sensors is impossible, there existed frequency-drift due to different topic publish rate of sensors. Rather than endow the camera pose of the nearest frame to the slave frames, we derive the accurate camera pose for slave frames in a spatially variant way. To enable accurate utilizing of observations from multiple cameras, a frequency-drift compensated strategy is proposed to eliminate the discrepancy of the corresponding frames. In our frequency-drift compensated strategy, for each slave frame, we make a hypothesis that there exist a corresponding reference frame F_{fic}^{i} named "fictitious frame" with the same timestamp and the slave frames can be mapped to the timeline of the references sensor. Therefore, for each slave frame, two adjacent frames F_{fw}^{i}, F_{bw}^{i} can be found. As shown in Figure 3, the red dots are the reference frames, the blue dots are the slave frames and the yellow are the fictitious frames. For each slave frame, the corresponding adjacent frames are listed in the table. A linear basis is imposed on the translation and rotation to recover the accurate camera pose of the fictitious frame. A scale parameter S is first computed based the time interval between the fictitious frame and two adjacent frames in Equation(4) and the camera positon of the fictitious frame can be calculated as Equation(5).

$$\mathrm{S} = \frac{(t^{fic} - t^{fw})}{(t^{bw} - t^{fw})} \tag{4}$$

$$\begin{aligned} x^{fic} &= x^{min} + \left[abs\left(x^{fw} - x^{bw}\right) * S\right] \\ y^{fic} &= y^{min} + \left[abs\left(y^{fw} - y^{bw}\right) * S\right] \\ z^{fic} &= z^{min} + \left[abs\left(z^{fw} - z^{bw}\right) * S\right] \end{aligned} \tag{5}$$

Where t^{fw}, t^{bw} are the timestamp of two adjacent reference frames respectively. t^{fic} is the timestamp of the fictitious frame, which is equal to the timestamp of the current slave frame. (x^{fw}, y^{fw}, z^{fw}) and (x^{bw}, y^{bw}, z^{bw}) are the camera position of two adjacent reference frames. $(x^{fic}, y^{fic}, z^{fic})$ are the derived camera position of the fictitious frame.

While trilinear interpolation is used to interpolate translation quantities, rotations have to be interpolated over the sphere to achieve constant-speed motion. This is achieved by the slerp operation.

$$\mathrm{slerp}\left(\mathrm{S}, r^{fw}, r^{bw}\right) = \frac{\sin\left((1-S)\alpha\right)}{\sin(\alpha)} r^{fw} + \frac{\sin(S\alpha)}{\sin(\alpha)} r^{bw} \quad , with\ S \in [0,1] \tag{6}$$

Which linearly interpolates between two quaternions r^{fw}, r^{bw} respectively, and where $\cos(\alpha) = r^{fw} \cdot r^{bw}$. More information on the slerp operation is descripted in (Shoemake, 1985).

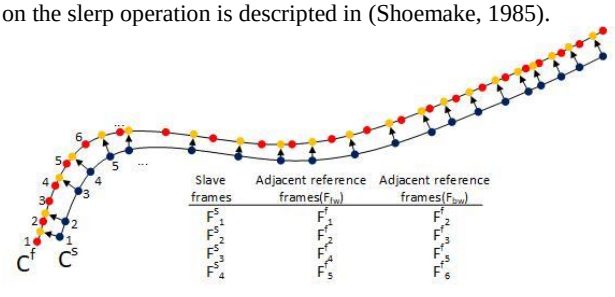

Figure 3. Fd-C strategy

3.2 Multiply camera tracking based on the drift-compensated frames

The implementation of our SLAM system is based on the ORB-SLAM framework. It mainly consists of two separate threads, in which images from multiple sensors are utilised: The first thread are responsible for reference camera tracking by observations from the adjacent frames. The second thread integrates the slave frame for pose refining. The transformation between the slave frame and the reference frames are obtained by the strategy in Section 3.1. As shown in Figure 4, in the first thread, the initial pose are derived with the reference camera. The second thread detected the adjacent frames for the slave frame continuously, the rigid transformation T^{rs} between the slave frame and its corresponding reference frame is calculated. Therefore we make a combination for each reference frame and its closest slave frame. The adjacent combinations are then used for pose refining.

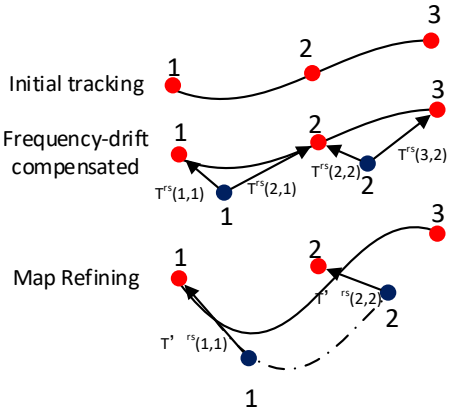

Figure 4. Framework for multiple camera tracking

After the Fd-C procedure, each combination are consisted of one reference frame F_r^i and one closest slave frame F_s^j, which's relation can be represented by a rigid transformation $T^{rs}(i,j)$. As the initial poses are derived from a single camera, the observation from the slave frames are underused. The frame combinations are used for pose refining. For the adjacent combination, two set of feature points $Pr_{fw} = \{Pr_{fw}^1, Pr_{fw}^2, \dots, Pr_{fw}^N\}$ and $Pr_{bw} = \{Pr_{bw}^1, Pr_{bw}^2, \dots, Pr_{bw}^N\}$ can be obtained from the reference frames and two set of feature points $Ps_{fw} = \{Ps_{fw}^1, Ps_{fw}^2, \dots, Ps_{fw}^M\}$ and $Ps_{bw} = \{Ps_{bw}^1, Ps_{bw}^2, \dots, Ps_{bw}^M\}$ can be obtained from the slave frames. The relations between Pr_{fw} and Ps_{fw}, Pr_{bw} and Ps_{bw} can be represented as following:

$$\begin{aligned} Pr_{fw} &= T_f^{rs} * Ps_{fw} \\ Pr_{bw} &= T_b^{rs} * Ps_{bw} \end{aligned} \tag{7}$$

In which T_f^{rs} and T_b^{rs} are the rigid transformation of the slave frame and the corresponding reference frame, which derived with the transformation from frequency-drift compensating and the external orientation parameters.

Thus, the multiply camera tracking means solving the following minimizations problem:

$$\begin{aligned} &\min_{P,R,t} \left(\sum_{i=0,j=0}^{M+N} \left\| C^r Pr_{fw}^i - \left[pr_{fw}^i, 1\right]^T \right\|^2 + \right. \\ &\left\| C^r (RPr_{fw}^i + t) - \left[pr_{bw}^i, 1\right]^T \right\|^2 + \left\| C^r Ps_{fw}^j - \right. \\ &\left. \left[ps_{fw}^j, 1\right]^T \right\|^2 + \left\| C^r T_f^{rs-1} (RPr_{fw}^i + t) T_b^{rs} - \right. \\ &\left. \left. \left[pr_{bw}^i, 1\right]^T \right\|^2 \right) \end{aligned} \tag{8}$$

Where pr^i_{fw}, pr^i_{bw}, ps^j_{fw}, pr^i_{bw} are the corresponding image points of the points $Pr_{fw}, Pr_{bw}, Ps_{fw}, Ps_{bw}$ respectively. $[R, t]$ is the refined pose by multiple camera tracking method.
This can be solved by iterations of nonlinear least squares. The pose updates of the adjacent frame would be refined and updated.

4. EXPERIMENTS AND ANALYSIS

In order to evaluate the accuracy of the camera trajectory, we use the trajectory of the GeoSLAM ZEB-REVO system(Cadge, 2016). As shown in Figure 5, to make sure the consistency of the trajectory between the Kinects system and ZEB-REVO system, ZEB-REVO device is fixed on the multiple Kinects system. The time-drift between two devices is compensated to insure the same initial position. To verify the performance of the proposed multiple RGB-D mapping solution, two set of datasets are collected. With the assumption of fix transformation between the position of Kinect and the laser at the same timestamp. The RMSE of the trajectory error of Kinect system can be calculated by Equation (9):

$$RMSE_X = \sqrt{\frac{\sum_{i=0}^{N}[abs(X_i^k - X_i^l) - DIS_X]^2}{N}}$$
$$RMSE_Y = \sqrt{\frac{\sum_{i=0}^{N}[abs(Y_i^k - Y_i^l) - DIS_Y]^2}{N}}$$
$$RMSE_Z = \sqrt{\frac{\sum_{i=0}^{N}[abs(Z_i^k - Z_i^l) - DIS_Z]^2}{N}} \quad (9)$$
$$\text{RMSE} = \sqrt{\frac{\sum_{i=0}^{N}[DIS_i - DIS]^2}{N}}$$

In which, $DIS_i = \sqrt{(X_i^k - X_i^k)^2 + (Y_i^k - Y_i^l)^2 + (Z_i^k - Z_i^l)^2}$, and $DIS = \sqrt{DIS_X{}^2 + DIS_Y{}^2 + DIS_Y{}^2}$

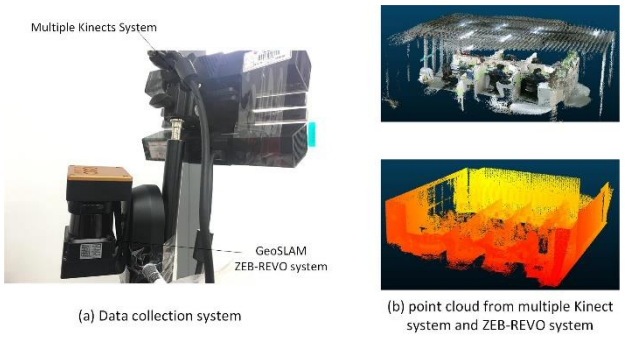

Figure 5. Data collection system and the point cloud of ZEB-REVO and RGB-D system

The absolute trajectory error (ATE) is calculated by Equation (6). It calculates the RMSE of the Euclidean distances between the camera trajectory and the timestamp associated ground truth, and it is used to evaluate the procedure's accuracy.
As shown in Table 1, it lists the absolute trajectory error of the results with single sensor and with multiple sensor. The results shows that the proposed Fd-C multiple RGB-D SLAM method achieve better in all conditions comparing with the results from single sensor. This can be explained by that more reliable visual features can be obtained using the dual Kinects. Lacking of reliable visual features in the single-Kinect system could introduce larger pose drifts, which will be accumulated during the whole operation.

Table 1. Comparisons of the absolute trajectory error for incremental registration of RGB-D sequences with single Kinect and dual Kinects

Dataset	Sensor count	RMSE.X (m)	RMSE.Y (m)	RMSE.Z (m)	RMSE (m)
Set1	1	0.119	0.145	0.081	0.205
	2	0.113	0.143	0.069	0.194
Set2	1	0.223	0.217	0.121	0.334
	2	0.185	0.193	0.092	0.283

Figure 6 and Figure 7 show the point cloud build by the system without Fd-C and with Fd-C. In Figure 6(a), the point cloud are obtained by directly merging the point cloud from upward sensor and downward sensor with the calibrated EoPs. In Figure 6(b), the point clouds are obtained by the tracking results with the Fd-C solution. Significant discrepancy appeared in the point cloud before Fd-C, which can be found from components structure in different views. Rather than large discrepancy in point cloud, our proposed multiple RGB-D mapping method produces a better results. The inconsistency of tracking results between different sensor was eliminated and generated more accurate point cloud. As expected, in Figure 7, the mapping results in dataset 2 shows a similar results. The proposed multiple RGB-D mapping method achieve more accurate results and again verifying its effectiveness.

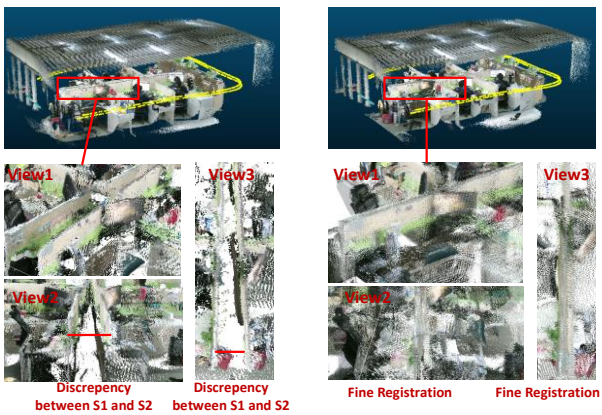

Figure 6. Comparison of the dense point cloud between before and after Fd-C of Dataset1

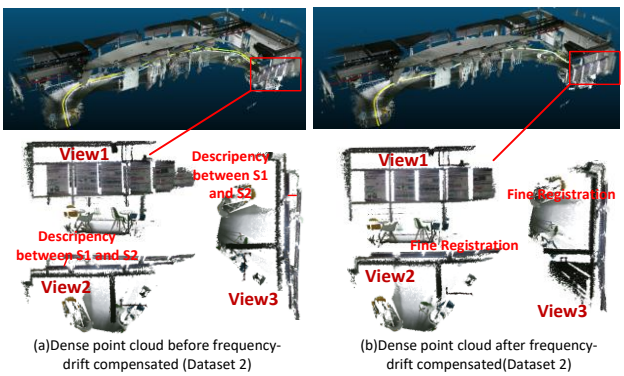

Figure 7. Comparison of the dense point cloud between before and after Fd-C of Dataset2

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W5-293-2019 | © Authors 2019. CC BY 4.0 License.

5. CONCLUSION

In this paper, we proposed using multiple RGB-D cameras in visual SLAM for better pose tracking performance and more detailed indoor environment mapping. We proposed a frequency-drift compensated method to eliminate the influence of time-drift between different camera during multiply camera motion tracking. Detailed mathematical analysis is presented to explain how to fuse all measurements from multiple camera for pose tracking. Through theoretical analysis and experimental validation, we conclude that the dual-Kinect mapping system is able to achieve better pose performance than single Kinect, and the proposed Fd-C multiply RGB-D mapping solution can eliminate the inconsistency between different sensors and produce better mapping results.

REFERENCES

Cadge, S., 2016. Welcome to the ZEB REVOlution. GEOmedia 20.

Chen, C., Yang, B., Song, S., Tian, M., Li, J., Dai, W., Fang, L., 2018. Calibrate Multiple Consumer RGB-D Cameras for Low-Cost and Efficient 3D Indoor Mapping. Remote Sensing 10, 328.

Darwish, W., Li, W., Tang, S., Wu, B., Chen, W., 2019. A Robust Calibration Method for Consumer Grade RGB-D Sensors for Precise Indoor Reconstruction (May 2018). IEEE Access, 1-1.

Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., Burgard, W., 2012. An evaluation of the RGB-D SLAM system, Robotics and Automation (ICRA), 2012 IEEE International Conference on, pp. 1691-1696.

Kerl, C., Sturm, J., Cremers, D., 2013. Dense visual SLAM for RGB-D cameras, Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on, pp. 2100-2106.

Mur-Artal, R., Tardos, J.D., 2017. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. IEEE Transactions on Robotics PP, 1-8.

Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A., 2011. KinectFusion: Real-time dense surface mapping and tracking, Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on, pp. 127-136.

Shoemake, K., 1985. Animating rotation with quaternion curves, ACM SIGGRAPH computer graphics. ACM, pp. 245-254.
Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D., 2012. A benchmark for the evaluation of RGB-D SLAM systems, 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 573-580.

Tang, S., Chen, W., Wang, W., Li, X., Darwish, W., Li, W., Huang, Z., Hu, H., Guo, R., 2018. Geometric Integration of Hybrid Correspondences for RGB-D Unidirectional Tracking. Sensors (Basel, Switzerland) 18, 1385.

Tang, S., Zhu, Q., Chen, W., Darwish, W., Wu, B., Hu, H., Chen, M., 2016. Enhanced RGB-D Mapping Method for Detailed 3D Indoor and Outdoor Modeling. Sensors 16, 1589.

Thrun, S., 2002. Probabilistic robotics. Commun. ACM 45, 52-57.

Whelan, T., Kaess, M., Fallon, M., Johannsson, H., Leonard, J., McDonald, J., 2012. Kintinuous: Spatially extended kinectfusion.

Yang, S., Yi, X., Wang, Z., Wang, Y., Yang, X., 2015. Visual SLAM using multiple RGB-D cameras, Robotics and Biomimetics (ROBIO), 2015 IEEE International Conference on. IEEE, pp. 1389-1395.