ICAS 2016
30th Congress of the International Council of the Aeronautical Sciences
DCC, Daejeon, Korea ; September 25-30, 2016

DEPTH IMAGE BASED DIRECT SLAM FOR SMALL UAVS

SungYeon Park*, David Hyunchul Shim*
***Department of Aerospace Engineering, KAIST, Daejeon, South Korea**

Keywords: *Real-time, Depth camera, SLAM, featureless method, indoor navigation*

Abstract

We propose a real-time keyframe-based direct (featureless) SLAM method utilizing a single depth image stream for small UAVs. The real-time performance of tracking and mapping is guaranteed on an embedded computer that has less computational resource. The camera motion is estimated by the direct depth image alignment based on keyframes in contrast with previous point-cloud approaches. The tracking speed is greatly improved by the approach and the decrease in the number of pixels by reduction of the image resolution and the intensity gradient based pixel sampling method. A 3D consistent map is reconstructed through the pose-graph optimization. The proposed framework is implemented on an actual UAV platform with a small embedded computer, and is compared with the ground-truth data for validation. It is proven to be effective for UAVs that mainly move in longitudinal direction with rotation.

1 Introduction

The navigation system is crucial for unmanned aerial vehicle (UAV) operations because most decisions are made based on the UAV location. Simultaneous localization and mapping (SLAM) algorithms are being developed as a complement of GPS systems whose utility range is limited to outdoor missions only. SLAM algorithms can help UAVs to operate in GPS-disabled areas and obtain geographic information. It has been one of the most frequently researched topics during the last decade, but the high computational load of SLAM algorithms is being an obstacle to actual implementation. Therefore, we aim to develop a fast and appropriate SLAM algorithm for indoor UAV operations.

The computational resource of small UAVs is strictly limited by payload and capacity of batteries. Recently, direct SLAM approaches have gained popularity since they use more information in an image compared to existing feature point approaches. Most state-of-art algorithms, however, cannot achieve their performance on a small embedded computer because their high computational complexity require a high performance CPU [8,13] or GPGPU acceleration [9,10,12]. It has been resolved by sending sensor data to a high performance ground station and receiving the navigation results [7]. These method is insufficient since the utility range is still limited by communication quality, and the quality is hardly stable when a UAV is in GPS-disabled areas. Therefore, an on-board real-time navigation system is positively necessary to be operated autonomously.

In this paper, a direct SLAM algorithm utilizing a single depth image stream is proposed, and the real-time performance is guaranteed on a small embedded computer. The direct image alignment is used to estimate the camera motion. It requires usually less computational resource than point-cloud approaches, and besides, the reduction of image resolution and the intensity gradient based pixel sampling method decrease the number of pixels to be calculated. In addition, inherent depth information in images makes the depth map estimation unnecessary in contrast with other SLAM methods using RGB cameras.

The system is appropriate for small UAVs. It is real-time guaranteed and costs less payload. It requires a single depth image sensor and a small embedded computer only. It is also effective for UAVs which mainly move in longitudinal direction with rotations, because the knowledge of depth information allows the system to be independent to the photo-consistent

assumption and the parallax. Monocular SLAM algorithms require the sensor to operate in continuous horizontal motions to obtain disparity in an image stream [2,4], and cause energy dissipation of the platform from the unnecessary movement. We integrate the sensor tracking system into LSD-SLAM framework [3] for the pose-graph optimization of a 3D consistent map. The framework is implemented on an actual UAV platform, and a real-world result for a trajectory which consists of forward and rotational motions is compared to the ground-truth data to validate the system.

The paper is organized as follows. Chapter 2 gives an explanation of the overall algorithm. The reduction of image resolution for decrease in number of pixels and abatement of the edge ambiguity is introduced in section 2.1. Section 2.2 explains the pixel sampling algorithm for speed up of error convergence and suitability to low depth feature environments. Section 2.3 describes the camera tracking algorithm using the depth image alignment. The map reconstruction process is introduced in Section 2.4. Chapter 3 presents experimental results under an actual flight condition. Lastly, achievements and future research directions are discussed in conclusion.

2 Depth Image based Direct SLAM

This Chapter introduces key concepts of the proposed algorithm. Figure 1 depicts the overall pipeline. The image stream contains depth information as pixel intensity. The first image is selected as a keyframe, and following images are used to estimate the sensor movement by the image alignment with the keyframe. When the latest image has much portion that do not overlap with the existing keyframe or the camera moves far from the previous keyframe location, the image is selected as a new keyframe for the following estimation process. The system keeps the history of keyframes and each element contains a depth map and positional relationship with the previous ones which are used for the pose-graph optimization to build a 3D map.

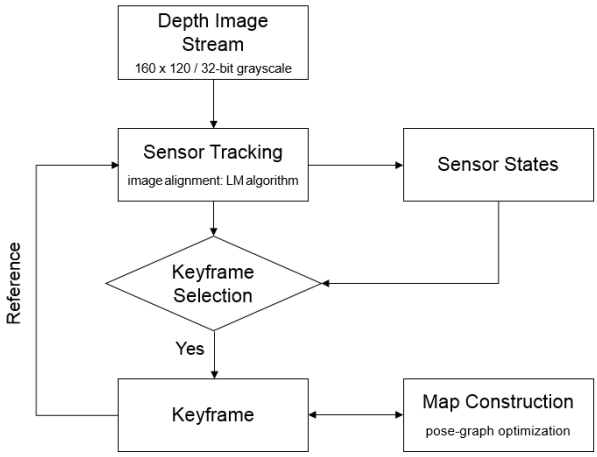

Fig. 1. Overall pipeline of the SLAM algorithm

2.1 Reduction of the Image Resolution

The sensor gives an image stream with VGA resolution. This huge number of pixels is excessive to estimate a camera movement and slows down the convergence speed [11]. The resolution is reduced to a sixteenth of an image, i.e. 160×120 resolution with 19,200 pixels. The representative value is found by averaging values of the pixels which are larger than five. The part of the compressed image is selected by the pixel sampling method and used to estimate the sensor movement.

Some negative features of the depth image stream such as the unmeasured (black) area and jagged object boundaries are caused by optical characteristics of the measurement method. The reduction not only improves the real-time performance but inhibits effects of these characteristics. As depicted in Figure 2, the black area is shrunk and the edges of objects become smooth during the process. Smaller resolutions should be avoided because the excessively blurred boundary causes the geometrical distortion of the environment.

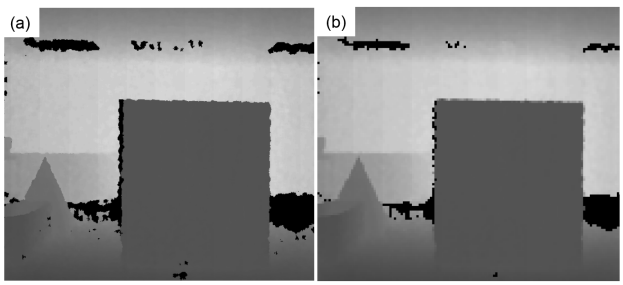

Fig. 2. (a) Original resolution, (b) Downsized image

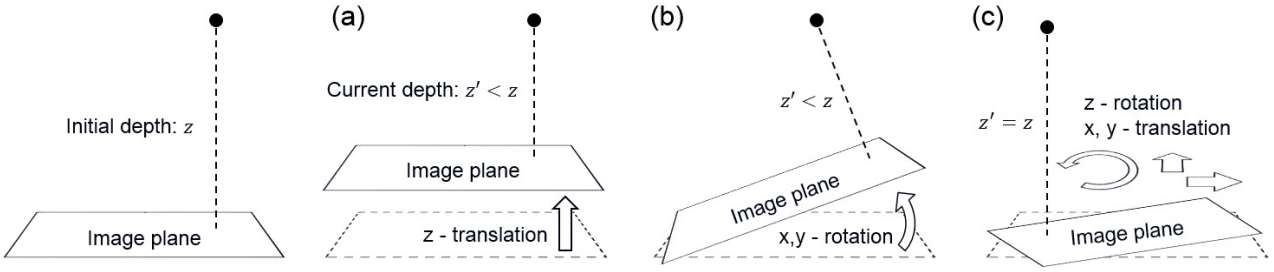

Fig. 3. (a)-(b) Transformations constrained by depth value, (c) Unconstrained transformations

2.2 Pixel Sampling Method based on Intensity Gradients

The overabundant pixels squander the computational resource of the platform, and some of them are not much useful to the image alignment process, e.g. pixels with zero gradients.

Because of the textureless characteristic of a depth image, a zero photometric error does not mean the correct alignment unlike in texture-rich images. As shown in Figure 3, when the sensor is facing the z axis, translation along z axis and the rotations about x and y axes change the depth value, but the value is not changed by the rotation about z axis and the translation along x and y axes [6]. If there is no gradient between adjacent pixels, the error will be still zero after planar motions of the sensor. The pixels with zero gradients exert a constraint on movements of only 3 DOF of the sensor.

Thus the pixels with non-zero gradients should be preferentially selected to estimate all 6 DOF movements of the sensor. The pixels with zero gradients and nonzero errors are uniformly sampled to keep the total number of pixels at least three thousand [5]. Inappropriate pixels are excluded for the further optimization process.

Table 1. Definition of the pixel sampling method

Condition	*Action*
The pixel is in unmeasured area	Excluded
Both of error and gradient values are zero	Excluded
The gradient value is zero	Uniformly sampled
Elsewhere (nonzero gradient)	Included

2.3 Image Alignment on SE(3)

The latest image frame I_{new} is aligned with the current keyframe I_{kf} to estimate camera movements. The alignment is done by minimization of the photometric error. The error of each sampled pixel is defined along the projective matching scheme (see Fig.4) and is given as:

$$r_i(\xi) = I_{new}\big(\omega(X_i,\xi)\big) - I_{kf}(\omega(X_i,\xi),\xi) \qquad (1)$$

Where the X_i is each pixel coordinate on the keyframe image, $I(X)$ or $I(X,\xi)$ mean the pixel value at coordinate X on the image I warped by the estimated camera movement $\xi \in se(3)$, and $\omega(X,\xi)$ projects a coordinate X from the reference image into the current image with ξ.

The error is iteratively minimized by the weighted Levenberg-Marquardt method in a left compositional formulation [1].

$$\delta\xi_k = -\big(J^T W J + \mu_k diag(J^T W J)\big)^{-1} J^T W r(\xi_k) \qquad (2)$$

$$\xi_{k+1} = \delta\xi_k \circ \xi_k \qquad (3)$$

The increment of the k[th] iteration $\delta\xi_k$ is computed by the equation (2) using the k[th] error residual $r(\xi_k)$ and the Jacobian matrix. The damping factor μ_k and the weight W are given for each iteration. The operator $\circ$ is defined as:

$$\xi_i \circ \xi_j \coloneqq \log\big(\exp(\xi_i)\cdot\exp(\xi_j)\big) \qquad [1] \quad (4)$$

The target error function is defined as a weighted sum of the photometric errors.

$$E(\xi) = \sum_{I_{kf}} w_i r_i(\xi) \qquad (5)$$

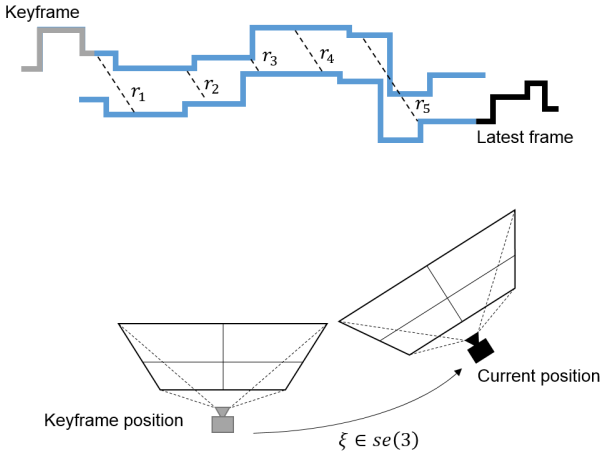

Fig. 4. Definition of error using the projective matching. The blue parts show overlapped portion between the latest frame and the keyframe.

A Jacobian matrix is the rate of change of each error metric with respect to the infinitesimal sensor movement ϵ.

$$J = \frac{\partial r(\epsilon \circ \xi_k)}{\partial \epsilon}\bigg|_{\epsilon=0} \tag{6}$$

The Jacobian matrix is found by addition of two matrices. One is the matrix from an image transformation in relation to the camera movement, and the other is from a direct change of the pixel value, i.e. objects become dark when the sensor get closer. The former is expressed with the gradient between adjacent pixels and has constraints for all 6 DOF. The direct change constrains only three types of movements as shown in Figure 3.

$$J = J_{image\ warp} + J_{direct\ change} \tag{7}$$

2.4 3D Map Reconstruction

The certain frames among an image stream are selected as keyframes and are used to reconstruct a 3D consistent map using their own depth maps and camera-to-world models. There are alignment errors between each keyframe because of the incompleteness of the tracking system. Thus, the pose-graph optimization is employed to readjust each depth map and remove the drift.

In contrast to monocular SLAM algorithm using a color image stream, a depth map of a depth image frame is directly obtained from the image when the keyframe is created. It is immediately used for the optimization process and abates drift accumulation, then the commotional position readjustment is diminished during the process.

3 Experimental Results

The flight experiment using an actual UAV platform is conducted to verify the proposed SLAM algorithm. The UAV moves along the predefined rectangular trajectory and conducts hover flight. The system is verified by a trajectory comparison between the real-time estimated path, the optimized trajectory, and the ground-truth data from the motion capture system. The final position of readjusted keyframes constitutes the optimized trajectory.

3.1 Experimental Setup

The constructed platform shown in Figure 5 has a width and length of 60 cm, a height of 25 cm, and a weight of 2 kg. It includes a LiPO battery with a weight of 455 g. Occipital Structure sensor equipped in a forward facing position is used as a depth sensor. It gives a depth image stream with VGA resolution (640 × 480) at 30 Hz and has a 58° horizontal FoV, a 45° vertical FoV, and a weight of 99 g. It can be lighter when the built-in battery is removed. It is calibrated with a simple pinhole camera model, and it is assumed that there is negligible distortion. The system is executed on a Samsung Exynos 5422 on ODROID-XU4 with ROS environment. The estimation result is transmitted to the flight control computer and the ground control computer via UDP communication. The ROS package of the LSD-SLAM framework [3] is used for map visualization. The ground-truth trajectory is measured by a motion capture system called Eagle Digital Real-time System which covers an area of 5 × 5 × 3 m. The platform moves along a rectangular path and operates only pitch and yaw motions to mimic common UAV movements which are mainly composed by longitudinal movements with rotations. Some objects are placed in the space to generate sufficient depth features in the field of view.

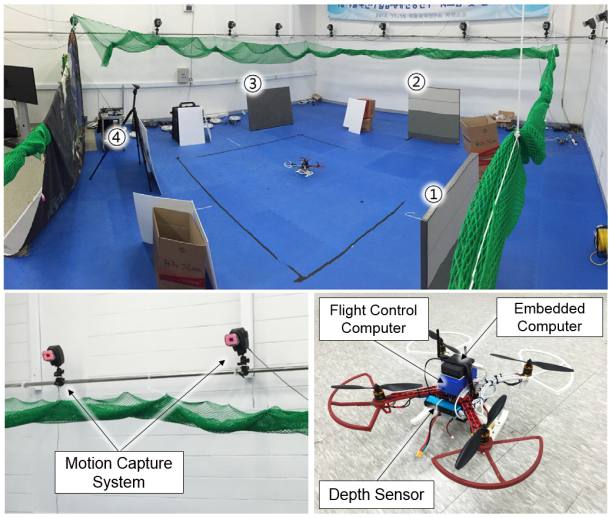

Fig. 5. Experimental setup. Objects for depth features, the motion capture system, and a UAV platform.

3.2 Results

In order to be used as a navigation system of a UAV, the translational root mean square error (RMSE) of the real-time position estimation and the calculation rate should be observed. The calculation time is obtained by averaging the time taken for two hundred images during the tracking and the mapping. The optimized trajectory and the map are not the essential information for a UAV but worth as a flight log and a geographical modeling of the nearby environment.

The characteristics of the platform causes difficulties in tracking depth images. The platform vibrates and fluctuates in all directions during the flight, and the fast movement increases the inaccuracy of the tracking performance. Especially, the measurement noise of the sensor is about 4 cm under a usual situation, i.e. the objects are 3 – 4 m far from the sensor. It makes small objects undetectable and decreases the accuracy of the tracking and the optimization process by the distortion of objects. The effects from these issues should be considered cautiously.

3.2.1 Loop-closure Test

The loop-closure must be detected to enable the global localization in a given environment. This experiment is conducted to check whether the system detects the loop-closure when the sensor

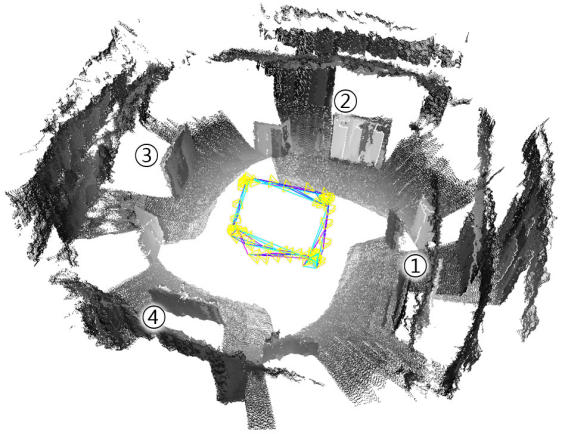

Fig. 6. Reconstructed 3D map for the loop-closure test

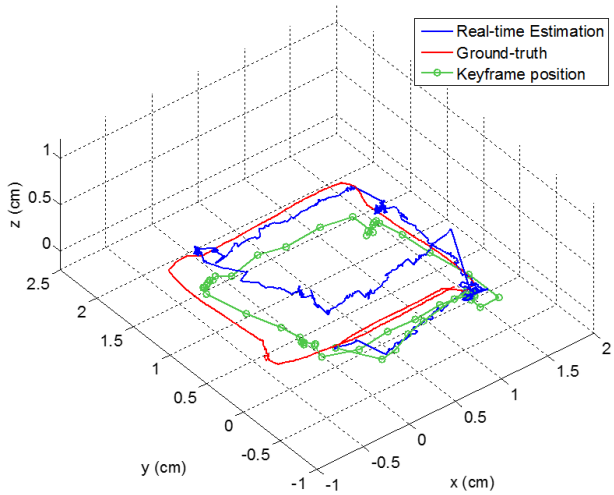

Fig. 7. Resulting trajectories

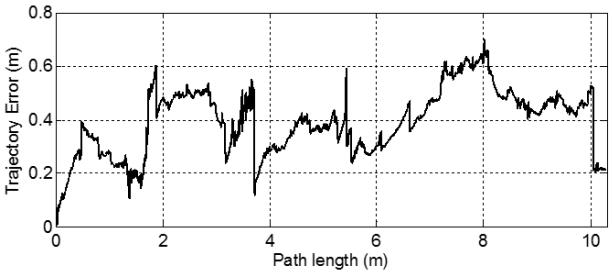

Fig. 8. Real-time estimation error due to path length

finished a cycle. The reconstructed 3D map and resulting trajectories are depicted in Figure 6, 7. The loop-closure was detected at 10.04 m, and the estimation error is noticeably decreased after the detection. The resulting parameters are listed in Table 2. The position of the sensor is stably estimated, but the error is diverges gradually in z direction. The major cause is that the system hardly estimates the exact rotation angle when the sensor conducts fast ninety degree rotations which are challenging for the system. The RMSE of the optimized trajectory is dropped by half in contrast to the real-time estimation.

Fig. 9. Some objects are placed for depth features

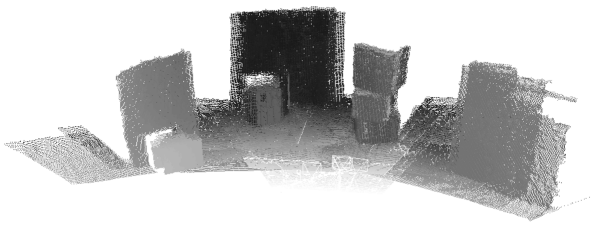

Fig. 10. Reconstructed 3D map for the flight test

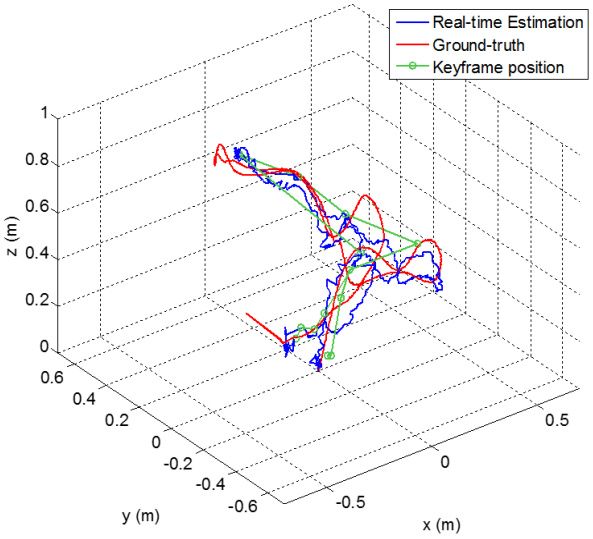

Fig. 11. Resulting trajectories

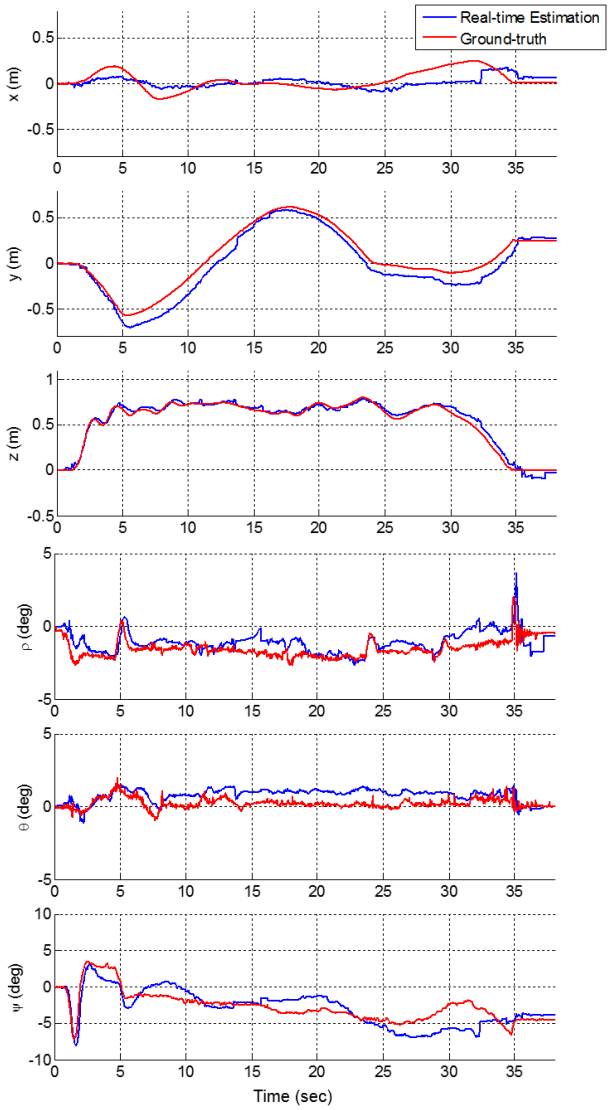

Fig. 12. Comparison of the trajectories about each axis

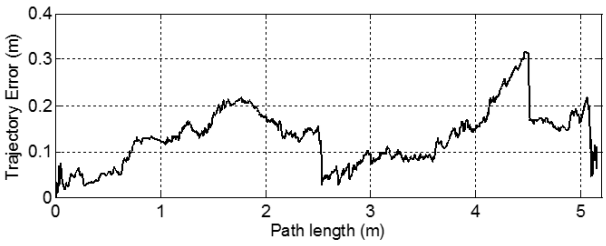

Fig. 13. Real-time estimation error due to path length

3.2.2 Flight Test

Both of the state estimation and the map reconstruction are fairly performed under the actual flight condition, but the estimation of the horizontal motion along the x axis is relatively insensitive. It might be caused by the local minima issue. The planar movements tend to be misconceived as rotation because relatively small rotation angle covers the change of view. It causes the spherical distortion of space. The high sensitivity of the heading angle estimation is caused by the same reason. The RMSE of the real-time trajectory estimation is 0.14 m. The roll and pitch attitudes are estimated with an RMSE of 0.7 °, and the heading angle is estimated with an RMSE of 1.6 °. Other resulting parameters are listed in Table 2.

Table 2. Experimental results for the experiments

		LC	*Flight*
RMSE (m)	*Real-time estimation before loop-closure*	0.398	0.144
	Real-time estimation after loop-closure	0.215	-
	Optimized trajectory	0.184	0.053
Calculation time (ms/frame)		23.6	22.8
Flight duration (sec)		122.0	38.1
Path length (m)		10.27	5.16
Drift rate (m/m)		0.084	0.012

The system is fairly robust to forward and rotational motions, i.e. there is no tracking instability during the motions. An image frame can be calculated with a rate of 40 Hz. It is much faster than other featureless SLAM methods, e.g. LSD-SLAM has a rate of 8 – 10 Hz on the embedded computer. The optimized trajectory from the system is closer to the ground-truth, and the 3D map describes the geometrical environment of the space.

4 Conclusion

We proposed a keyframe-based featureless light-weight SLAM utilizing a single depth image stream. It is appropriate to small UAVs because the depth sensor is light enough and the computational load of the proposed algorithm is reasonable for the limited computational resource of the onboard computer that can be equipped on typical small UAVs. The real-time performance is achieved by immediate acquisition of depth maps, the direct image alignment, and the decrease of pixels. The commotional adjustment of the position estimation are abated by immediate map optimization of keyframes. The system reconstructs a 3D consistent map and enables the global localization by detecting the loop closure. The performance has been evaluated under the actual flight condition with the UAV platform. It can be useful for UAVs which of the computational resource is highly limited or for certain situations such as textureless and dark environments.

There are many directions for the future research. First, the overall performance can be improved by solving the local minima issue. A sensor with a broader FoV can abate the spherical distortion of space because the geometric relationship between objects become clear if the objects are observed in one image. Second, the precise calibration of the depth sensor can help the system estimate the location more accurately. The camera model is currently based on the simple pinhole camera model and assumed that there is negligible distortion in an image. Finally, the update rate is excessively fast compared to the rate of 30 Hz of the sensor. The system can be improved with additional algorithms to supplement the sensor tracking or the depth map refinement process.

Acknowledgement

This research was supported by the KAIST Institute Research Fund (grant No. N10150030).

References

[1] Caruso D, Engel J and Cremers D. Large-scale direct SLAM for omnidirectional cameras. *International Conference on Intelligent Robots and Systems (IROS),* Hamburg, Germany, pp. 141-148, 2015.

[2] Daftry S, Dey D, Sandhawalia H, Zeng S, Bagnell J A and Hebert M. Semi-Dense Visual Odometry for Monocular Navigation in Cluttered Environment. IEEE International Conference on Robotics and Automation (ICRA), Seattle, Washington, 2015.

[3] Engel J, Schöps T and Cremers D. LSD-SLAM: Large-Scale Direct Monocular SLAM. *In European Conference on Computer Vision (ECCV)*, Zürich, Switzerland, pp. 834-849, 2014.

[4] Engel J, Sturm J and Cremers D. Semi-dense visual odometry for a monocular camera. *International Conference on Computer Vision (ICCV)*, Sydney, Australia, pp. 1449-1456, *2013*.

[5] Fang Z and Zhang L. Real-time and robust odometry estimation using depth camera for indoor micro aerial vehicle. *The 26th Chinese Control and Decision Conference (2014 CCDC)*, Changsha, Hunan, pp. 5254-5259, 2014.

[6] Gelfand N, Ikemoto L, Rusinkiewicz S and Levoy M. Geometrically stable sampling for the ICP algorithm. *Fourth International Conference on 3-D Digital Imaging and Modeling*, Banff, Alta, pp. 260-267, 2003.

[7] Huh S, Cho S and Shim D H. 3-D Indoor Navigation and Autonomous Flight of a Micro Aerial Vehicle

using a Low-cost LIDAR. *Journal of Korea Robotics Society*, pp. 154-159, 2014.
[8] Kerl C, Sturm J and Cremers D. Dense visual slam for rgb-d cameras. *International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, pp. 2100-2106, 2013.
[9] Newcombe R. A, Lovegrove S J and Davison A J. DTAM: Dense tracking and mapping in real-time. *International Conference on Computer Vision (ICCV)*, Barcelona, Spain, pp. 2320-2327, 2011.
[10] Newcombe R A, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison A J, Kohi P, Shotton J, Hodges S and Fitzgibbon A. KinectFusion: Real-time dense surface mapping and tracking. *International symposium on Mixed and augmented reality (ISMAR)*, Basel, Switzerland, pp. 127-136, 2011.
[11] Pomerleau F, Magnenat S, Colas F, Liu M and Siegwart R. Tracking a depth camera: Parameter exploration for fast ICP. *International Conference on Intelligent Robots and Systems*, San Francisco, CA, pp. 3824-3829, 2011.
[12] Whelan T, Kaess M, Fallon M, Johannsson H, Leonard J and McDonald J. "Kintinuous: Spatially extended kinectfusion." Technical report, 2012.
[13] Zhang J and Singh S. LOAM: Lidar odometry and mapping in real-time. *Robotics: Science and Systems Conference (RSS),* Berkeley, CA, pp. 109-111, 2014.

Contact Author Email Address

sungyeon_park@kaist.ac.kr
hcshim@kaist.ac.kr

Copyright Statement