Translational Motion Estimation Using Kinect

Muhammad Fuad
Faculty of Engineering, University of Trunojoyo Madura, Bangkalan, Indonesia

ibrahim.fuad@gmail.com, fuad@trunojoyo.ac.id

Abstract. Study on robot navigation has been progressing significantly. However, research of autonomous navigation is still lacking in environments with changing layout. For supporting robot navigation system, rotational and translational motion estimation need to be improved. Low cost sensor such as Kinect can be used as alternative to laser to percept changing in the vicinity of robot. There is still problem in developing motion estimation using Kinect. Narrowness in the sensor field of view limits its ability to track robot motion. This problem emerges specially in translational motion estimation. This paper aims to develop a simple approach based on RGB-D data for translational motion estimation of a Kinect sensor. RGB-D data consist of RGB and Depth images. In previous research, camera was positioned in the direction of longitudinal translational motion. This paper differs from previous works in terms of sensor setting that is placed in the direction of lateral translation movements. Point features are matched between pairs of RGB frame to get interesting pixels. Depth image provides distance information for these pixels in synchronization with feature tracking process. Law of cosines is then applied to these depth pixels to get position translation. Results show that our approach have decreased RMSE to 0.0382 meter better than previous works. In the future research, this translational motion estimation system will be implemented in mobile robot system to support navigation function.

1. Introduction
Autonomous navigation is one of important application in controls and robotics research domain. For supporting navigation function, video sensor should be used effectively as measurement unit to replace laser that has more weight and more expensive. Video and visual perception system plays central role in pose estimation and localization as part of navigation function to provide current position (x, y) and orientation (ψ) information of robot. This information, known as robot pose (x, y, ψ), have essential value not only for navigation but also for environmental mapping. By tracking the change of pose, one can estimate rotational and translational motion. This problem is well-known as motion estimation problem. A solution of motion estimation based on sequence image analysis on video sensor data is called visual motion estimation [1].

Pose estimation with combination of laser, conic mirror and a fish-eye camera was implemented in a crawling robot for inspecting a pipeline [2]. Motion estimation based on fish-eye images that can be applied for video surveillance and automotive was investigated in [3]. Pose estimation relied on 3D model was studied in [4] for augmented reality. Pose estimation to form error signals for controlling Unmanned Aerial Vehicle (UAV) in environment without GPS was explored in [5]. Pose estimation used single web camera was worked in [6]. Pose and motion estimation are beneficial not only for

1569 (2020) 032067 doi:10.1088/1742-6596/1569/3/032067
application in control and robotics but also advantageous for other fields such as video surveillance, automotive and augmented reality.

More cost has to be spent to improve precision in laser-based pose estimation. More computation time is needed when some sensors combined in order to deliver more data for estimation process. Model is not always available especially in environment with changing layouts. This paper aims to develop a simple approach to estimate motion by using relative low-cost sensor compare to laser. Kinect is chosen as sensor in this work because it has ability to supply both RGB images and depth (RGB-D) data. In order to complete our previous work on Kinect-based orientation estimation [7], this paper presents a new method to estimate translation movement based on RGB-D sensor. Our approach differs from previous researches in terms of Kinect sensor setting that is placed in the direction of lateral translational motion. This work contributes to present new sensor configuration and simple method to estimate translational motion. Pairs of RGB images are matched by means of point features. Depth data of these matched pairs are then be used to compute translation displacement by employing a triangulation approach. Our method has been implemented in Intel Core i3-2350M CPU 2.3GHz.

The rest of this paper is organized by using following structure. Section 2 contains review of some related works on pose and motion estimation. Section 3 describes our proposed methods. Section 4 discusses the result of our experiments. Section 5 concludes this paper.

2. Related Works
An approach to perform visual odometry and scene flow by means of RGB-D camera based on geometric cluster of scene segmentation was worked in [8]. An effort to separate estimation for rotational and translational camera motion by exploiting orthogonal planar structures was performed in [9]. The similar approach by utilized orthogonal planar features as in [9] was extended in [10] for improving RGB-D-based SLAM. An attempt to use Depth image only from RGB-D camera for decreasing drift of SLAM was presented in [11]. An integration of both depth and RGB images to widen range in distance measurement was investigated in [12]. A research to present a new data set of new generation of RGB-D camera was described in [13]. Their RGB-D data set was synchronized from two sensors. This data set was evaluated by using RGB-D visual odometry algorithm. Three models of point feature spatial uncertainty were introduced in [14]. The simplest model was used in idealized SLAM system.

A solution of motion estimation based on RGB-D sensor only that called RGB-D odometry investigated in [15]. Point and line features were fused to produce smaller motion estimate uncertainty. Their proposed method was called Point and Line Visual Odometry (PLVO). Speeded Up Robust Features (SURF) algorithm was used for detecting 2D point features from RGB images. While Mean – Standard deviation Line Descriptor (MSLD) was employed for line identification. For testing PLVO in real-world scenarios, a Kinect was hand-held in indoor environment. Camera orientation was aligned with direction of human walking motion. Trajectory Endpoint Drift (TED) was reported to reach value 1.93 meters for 45 m travel distance in classroom environment with constant lighting. This evaluation metric was describing global consistency by comparing ground truth and estimated translational motion. For describing error of translational drift, their method got 0.049 meter by using Root Mean Squared Error (RMSE) of Relative Pose Error (RPE) evaluation metric.

A fast visual odometry from RGB-D camera was presented in [16] by using an intensity-assisted Iterative Closest Point (ICP). Selection point selection was performed by using intensity residual, intensity gradient and depth gradient. Matching point was determined by considering intensity and geometric distance from nearby points in the image coordinate. Weighting of corresponding pairs was executed based on robust statistics. RMSE of translational drift was 0.0416 meter per second for experiment in room environment. All images and depth data were recorded in resolution 640 x 480.

An effort to decouple translational from rotational motion by means of lines and planes in Line and Plane based Visual Odometry (LPVO) was carried out in [17]. The average of translational RMSE was 0.04 meter. In the most part of their work, camera was positioned in the direction of longitudinal
translational motion. In this direction, LPVO got sufficient planes to extract information. Although LPVO only catched single plane, camera motion still stably tracked because of the information extracted from line in that plane. Some of other Visual Odometry (VO) algorithms fail to estimate when they were provided single plane only. The performance of LPVO inspired us to develop a new translational motion estimation based on single plane.

3. Method

Figure 1 shows our approach to develop translational motion estimation based on single plane of RGB-D images. Our proposed system consists of point feature detection, point feature matching, RANSAC-based motion estimation, pixel-wise distance estimation and triangulation-based translational motion estimation. Point features are detected in point feature detection process for each RGB frame. Point features from pairs of RGB color images are matched during the second process. the set of features matched are then fed into motion estimation based on RANSAC. On the other part, Depth images are extracted into distance information of each pixel. By using these distance information, Law of Cosines-based process is used to compute translational motion.

In order to support our algorithm, this paper proposes camera configuration in a mobile robot system such as illustrated in figure 2. Front camera is positioned at a standard direction to observe front view. In this direction, front camera will monitor more than one plane. At least there are four planes in front direction that consist of floor, ceiling, left and right wall. By positioning two dedicated cameras in left and right of mobile robot body, each camera will only perceive single plane. RGB and Depth images are provided in 640x480-pixel resolution. Both images are captured in 30 frame per second (fps).

3.1. Pixel-wise distance estimation

Based on our experiments, Kinect has limitation in distance that can be measured. Figure 3 shows that sensor cannot measure distance more than 6 meters. Objects in the corridor can be viewed in the RGB images but not in the depth images. Depth image displays distance objects in white color. On the other hand, distance less than 1 meter also cannot be sensed as displayed in figure 3. Pixel of objects at the wall that have near distance are displayed in white color as well. This research makes use Kinect that positioned in the side of mobile robot to capture depth information in the range between 1 meter to 6 meters.
3.2. Point Feature Detection
Visual feature is the most interesting part of image that has sufficient property for detection. Corner is a feature that can be used for this process. This work uses point feature extraction that has less complexity to reduce computation time. This feature is not robust and stable but it can be handled by providing distance per pixel that completing 3D information of feature. Corners are points that have two dominant direction in luminosity gradient in the nearby points. Harris and Features from Accelerated Segment Test (FAST) are two algorithms that used and compare in this work to extract point features from RGB image. Figure 4.(a) shows results of point features that have detected using Harris and FAST.

3.3. Point Feature Matching
Feature tracking is known to be crucial stage in visual odometry. Maximum cross correlation is used for point feature matching in this work. A pair of images is needed in this process. Point features that have been extracted from current frame are matched with features that have been detected from previous frame. Figure 4.(b) displays the results of matching process on Harris and FAST features. As consequences of larger number of features, matching process on Harris features generates more pairs than on FAST features.

3.4. RANSAC-based motion estimation
Random Sample Consensus (RANSAC) is applied in matched pairs of point features. This step is used to eliminate false point feature pairs (outlier). As consequences, this process results trusted point feature pairs (inlier). Figure 4.(c) presents RANSAC implementation on Harris and FAST matched features.

3.5. Law of Cosines-based translational motion estimation
The last process in translation motion estimation is displacement computation of trusted point feature pairs. Figure 5 illustrates our approach for computing distance of translation motion. On the top left, a Kinect camera that face to single plane is moved to the front at discrete time $k-1$. At this time, camera capture three features in RGB image $I_{k-1}$. Let us concentrate on square as representation of a point feature $p_{k-1}$ that has been detected at $k-1$. When camera arrive at time $k$ by translational motion, this square has been moved to another part of image $I_0$ as point feature $p_k$. Arrows in these figures represent pixel-wise distance $d_{p_{k-1}}$ and $d_{p_k}$ respectively. If we put together arrows and matched point features from both images then a triangle will be formed such as depicted in the bottom of figure 5. Angle between arrows $\alpha_{p_k}$ equals to travelled distance of translational motion of point feature $i_{p_k}$ that can be computed using law of cosines as follows

$$\hat{t}_{p_k} = \sqrt{d_{p_{k-1}}^2 + d_{p_k}^2 - 2d_{p_{k-1}}d_{p_k} \cos(\alpha_{p_k})}$$

(1)
Not every trusted point feature pairs have distance information. This situation occurs when point feature of RGB image is in the blind area of camera as described in section 3.1. Therefore, estimated translational motion should be computed by averaging magnitude of translational motion of each trusted point feature that has distance information such as expressed as follows

\[ \hat{t}_k = \frac{1}{n} \sum_{k=1}^{n} \hat{t}_{p_k}, \quad d_{p_{k-1}} > 0, \quad d_{p_k} > 0 \]  (2)

For measuring accuracy of our approach, RMSE is utilized by equation as follows

\[ RMSE_t_k = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (\hat{t}_k - t_k)^2} \]  (3)

4. Results and Discussions

Translational motion estimation system has been implemented in indoor environment. A living room that has dimension 3.5 meters squared is chosen as environment. The dimension is selected with respect to range of distance that can be captured by sensor such as explained in section 3.1. We evaluate our approach using combination natural and artificial lighting. Our approach is tested for translation motion estimation with actual distance from 0.2 to 0.9 meter. Table 1 shows RMSE in meter for point features detected by Harris compare to features identified by FAST. According to our experiment results, Harris features produce less error than FAST. Parameter settings in both algorithms affect the performance of feature detection process. Figure 6 shows a sample of comparison between Harris and FAST on translation motion estimation.

5. Conclusion

Translational motion estimation based on Kinect is successfully developed in this research. RMSE of translational drift of our approach is 0.0382 meter and 0.0767 meter for experiments using Harris and FAST respectively.

![Figure 5. Translation estimation using law of cosines](image)

![Figure 6. Translation motion estimation using Harris and FAST for 0.7 meter actual travelled distance is resulting 0.7334 meter and 0.7278 meter respectively](image)

| No. | Actual Distance | RMSE (meter) |
|-----|----------------|--------------|
|     |                | Harris       | FAST         |
| 1.  | 0.2            | 0.0120       | 0.2128       |
| 2.  | 0.3            | 0.0358       | 0.0272       |
| 3.  | 0.4            | 0.0219       | 0.1129       |
| 4.  | 0.5            | 0.0289       | 0.0322       |
| 5.  | 0.6            | 0.0705       | 0.0273       |
| 6.  | 0.7            | 0.0309       | 0.0536       |
| 7.  | 0.8            | 0.0374       | 0.0746       |
| 8.  | 0.9            | 0.0681       | 0.0731       |
|     | Average        | 0.0382       | 0.0767       |
References

[1] A. Kelly, Mobile Robotics: Mathematics, Models, and Methods, New York: Cambridge University Press, 2013.

[2] A. D. Tezerjani, M. Mehrandezh, R. Paranjape, "4-DOF Pose Estimation of a Pipe Crawling Robot using a Collimated Laser, a Conic Mirror, and a Fish-eye Camera," in Proceeding of Southwest Symposium on Image Analysis and Interpretation, San Diego, CA, USA, April 2014, pp. 45-48.

[3] A. Eichenseer, M. Batz, J. Seiler, and A. Kaup, “A Hybrid Motion Estimation Technique for Fish-eye Video Sequences Based on Equisolid Re-Projection,” in Proceeding of IEEE International Conference on Image Processing (ICIP), Quebec City, Canada, September 2015, pp. 3565-3569.

[4] E. Marchand, H. Uchiyama, F. Spindler, “Pose Estimation for Augmented Reality: A Hands-On Survey,” IEEE Transactions on Visualization and Computer Graphics, Vol. 22, Issue12, pp. 2633 - 2651, December 2016.

[5] H. Xie , K. H. Low, and Z. He, "Adaptive Visual Servoing of Unmanned Aerial Vehicles in GPS-Denied Environments", IEEE/ASME Transactions On Mechatronics, Vol. 22, No. 6, December 2017, pp. 2554-2563.

[6] K. Singhirunrusorn, F. Fahimi, R. Aygun, "Single-camera pose estimation using mirage", IET Computer Vision, Vol. 12 Issue 5, pp. 720-727, April 2018, pp. 720-727.

[7] M. Fuad, “Visual Odometry Using Kinect Sensor,” in Proceeding of National Seminar of Information Technology and Multimedia (SVASTIA), pp. B-23 - B-31, September 2013.

[8] M. Jaimez, C. Kerl, J. Gonzalez-Jimenez, D. Cremers, "Fast Odometry and Scene Flow from RGB-D Cameras based on Geometric Clustering”, in Proceeding of IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 3992-3999.

[9] P. Kim, B. Coltin, H. J. Kim, "Visual Odometry with Drift-Free Rotation Estimation Using Indoor Scene Regularities", in Proceeding of British Machine Vision Conference (BMVC), 2017, pp. 1-12.

[10] P. Kim, B. Coltin, H. J. Kim, "Linear RGB-D SLAM for Planar Environments", in Proceeding of Europe Conference on Computer Vision (ECCV), 2018, pp. 350-366.

[11] S. Zhao, Z. Fang, "Direct Depth SLAM: Sparse Geometric Feature Enhanced Direct Depth SLAM System for Low-Texture Environments", in Sensors, 2018, 18, 3339; doi:10.3390/s18103339.

[12] S. Tang, Q. Zhu, W. Chen, W. Darwish, B. Wu, H. Hu, M. Chen, "Enhanced RGB-D Mapping Method for Detailed 3D Indoor and Outdoor Modeling”, in Sensors, 2016, 16, 1589; doi:10.3390/s16101589.

[13] M. Kraft, M. Nowicki, A. Schmidt, M. Fularz, P. Skrzypczy´nski, "Toward evaluation of visual navigation algorithms on RGB-D data from the first- and second-generation Kinect", in Machine Vision and Applications, 2017, 28, pp.61–74, DOI 10.1007/s00138-016-0802-6.

[14] D. Belter, M. Nowicki, P. Skrzypczy´nski, "Modeling spatial uncertainty of point features in feature-based RGB-D SLAM", in Machine Vision and Applications, 2018, 29, pp.827–844, DOI 10.1007/s00138-018-0936-9.

[15] Y. Lu, D. Song, "Robust RGB-D Odometry Using Point and Line Features", in Proceeding of IEEE International Conference on Computer Vision (ICCV), pp. 3934 - 3942, Santiago, Chile, Dec. 2015.

[16] S. Li and D. Lee, "Fast Visual Odometry Using Intensity-Assisted Iterative Closest Point", IEEE Robotics And Automation Letters, Vol. 1, No. 2, pp. 992 - 999, July 2016.

[17] P. Kim, B. Coltin, H. J. Kim, "Low-Drift Visual Odometry in Structured Environments by Decoupling Rotational and Translational Motion”, in Proceeding of IEEE International Conference on Robotics and Automation (ICRA), pp. 7247 - 7253, Brisbane, Australia, May 21-25, 2018.