Depth-based Direct Visual Odometry for Stereo Camera

By Sangil LEE,[1)] Pyojin KIM[1)] Changhyeon KIM[1)] and H. jin KIM[1)]

[1)] *Mechanical & Aerospace Engineering and Automation Systems Research Institute, Department of Mechanical & Aerospace Engineering, Seoul National University, Seoul, South Korea*

In this paper, we propose a depth-based stereo visual odometry algorithm which tracks the camera pose using the intensity error of the pixel directly. By using a direct method, the proposed algorithm can improve the robustness to high frequency oscillation or blurring of visual input, thus making it possible to estimate the camera pose accurately. We calculate a depth image from stereo calibration and disparity, which makes it easy to construct a 3D representation of pixels. It is difficult for a monocular camera to calculate scale factor, whereas stereo can estimate scale factor easily by fixed baseline. In the research, we use a stereo camera rather than a monocular one in order to compensate the scale drift error. Additionally, a keyframe has a role as a reference image to estimate the relative motion of the camera resulting in reduction of drift error. Finally, we validate the performance of the proposed algorithm with the EuRoC MAV dataset which is a vision database measured on an unmanned aerial vehicle.

Key Words: navigation, visual odometry, ego-motion, stereo camera, direct method, keyframe

Nomenclature

$\mathbf{x}$	:	2D pixel point in image plane
$I(\cdot)$	:	image intensity matrix
$r(\cdot)$	:	residual for optimization problem
Subscripts		
i	:	i-th image
j	:	j-th pixel

1. Introduction

Visual odometry is a computer vision technology to recognize the ego-motion of a camera with image sequence including pixel's azimuth, elevation and intensity information in the camera coordinates.[1,2] Since the traditional odometry uses inertial measurement unit or motor encoder to estimate the trajectory of a robot, it results in an error accumulation because of the friction of wheel or non-ideal physics model. However, the method using visual information is robust to such issues. Visual odometry is broadly used for estimating an ego-motion reducing a drift error with a visual clue, while the existing inertial odometry shows a large drift error.

Visual odometry can be classified into: 1) stereo vs. monocular, 2) feature vs. direct.[2] 1) While stereo visual odometry estimates the position and orientation of the camera coordinate system with two images captured simultaneously by a stereo camera, monocular visual odometry uses only one image. Generally, stereo visual odometry is sufficiently useful indoors, since it can easily calculate a depth of each pixel and estimate the exact positioning scale factor due to a fixed baseline which refers to the relative translation and rotation between two cameras of the stereo rig.[3] 2) When comparing two contiguous images captured at different times, we have to determine criteria in order to calculate the transformation matrix which describes the motion between two cameras coordinates. While feature-based visual odometry minimizes reprojection error using bundle adjustment to estimate the motion of a camera,[4,5] the direct method minimizes the photometric error.[6,7] Feature-based visual odometry can estimate the ego-motion of a camera in real-time, but it has to extract robust, precise features. On the contrary, direct visual odometry can estimate the ego-motion robustly and precisely though it has extensive computation load.

In this paper, we propose a depth-based stereo visual odometry. The proposed algorithm estimates the ego-motion of a camera based on the depth information derived from stereo image pairs so that its performance can be improved. Also, a keyframe selection based on the 2D correlation can reduce a drift error in a static movement. For the quantitative and qualitative evaluation of the proposed algorithm with a well-known stereo visual odometry algorithm, we use various error measurement methods for comparison.

2. Method

Figure 1 shows the brief overview of the proposed stereo visual odometry algorithm. The stereo image pairs from the synchronized camera are fed into the proposed algorithm. Then, the 2D correlation between the current and keyframe image is calculated. If the correlation is smaller than a heuristic threshold, the current image is selected as a new keyframe. For the rectification of an image, stereo calibration can help to align two images horizontally and calculate a depth of each pixel. Using the depth image, the proposed algorithm estimates the ego-motion by minimizing the photometric error between keyframe and the current image based on the 3D representation of pixel with an optimization parameter, i.e. relative motion.

2.1. Keyframe selection

Unlike the simultaneous localization and mapping (SLAM) which could memorize the place which the robot visited before by storing feature points, visual odometry does not construct the 3D representation of an environment, thus still showing non-negligible drift error. This drift error is prominent in a naive visual odometry which estimates the ego-motion images consecutively. Even in the case where images show a static movement, randomized errors can be accumulated causing es-

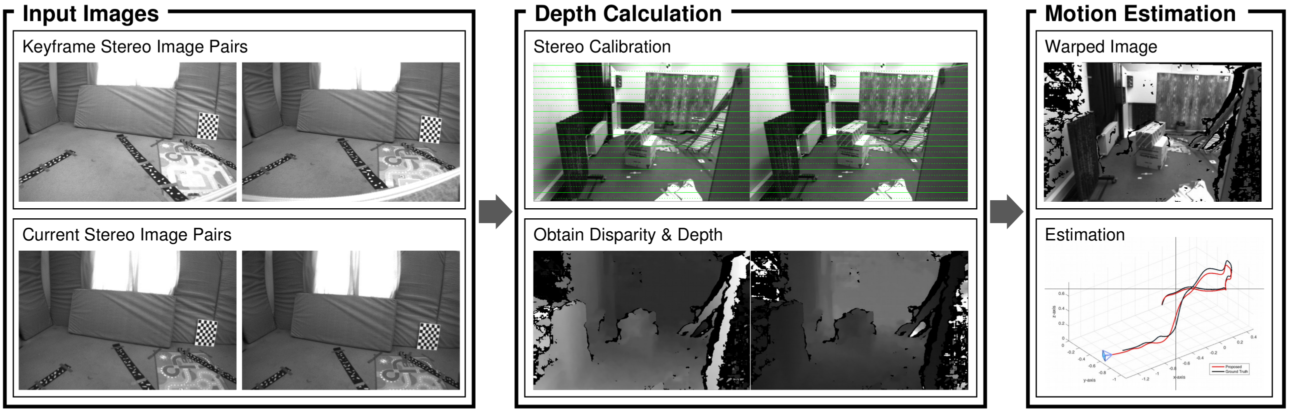

Fig. 1. **Overview of the proposed stereo visual odometry algorithm.** First, the correlation between the current and keyframe image is calculated. Second, through stereo calibration, we align two images horizontally and calculate the depth of each pixel. Based on the depth image, we estimate the ego-motion by minimizing the photometric error between keyframe and image warped with relative motion.

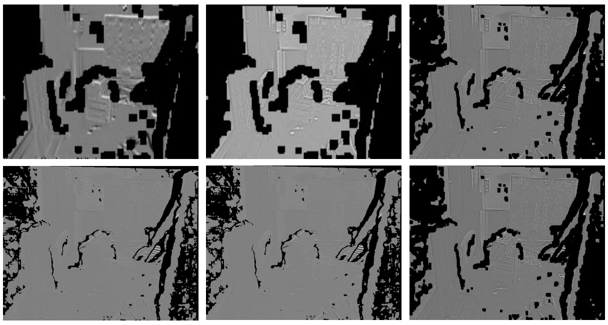

Fig. 2. **Optimization process.** Clockwise from top-left shows the process of optimization minimizing the photometric error.

timation to deviate from the true trajectory. Consequently, we use a keyframe selection and estimate the ego-motion between the current image and keyframe one. In order to select a proper keyframe, we use the 2D correlation coefficient algorithm. The output of the correlation algorithm varies from -1 to 1, and the higher output value, the better similarity between two images. The first stereo image should be a keyframe in the beginning, and the current image could be a new keyframe if the correlation between the current image and the previous keyframe is smaller than a heuristic threshold.

2.2. Depth calculation

A sequence of stereo images is fed into the visual odometry algorithm with known baseline, which is a relationship between two cameras of the stereo rig. By using the baseline, we know the 3D representation of the corresponding point of the two stereo image pairs. In detail, the stereo image pairs are rectified with the stereo calibration parameter. Then, the stereo calibration calculates the intrinsic matrix of the both cameras and align the both images horizontally. Thus, all the corresponding points in the rectified image become collinear. After all, by calculating the distance of the corresponding pixels in the same horizontal line, i.e. disparity,[8] we can estimate the depth of each pixel by means of an inverse of the disparity.

2.3. Direct ego-motion estimation

In direct approaches, the relative motion of a camera including translation and rotation is estimated iteratively, and optimized minimizing the photometric error simultaneously. The photometric error of j-th pixel in the i-th image sequence can be written as follows:

$$r_{ij}\left(\boldsymbol{\xi}_i\right) = I_i\left(w\left(\boldsymbol{\xi}_i, \mathbf{x}_j\right)\right) - I_{ref}\left(\mathbf{x}_j\right) \tag{1}$$

where $\boldsymbol{\xi}$ is the relative motion of the camera, and $w(\cdot)$ is a function which calculates the warped location of pixels in the keyframe image based on the given relative motion $\boldsymbol{\xi}$. The optimal relative motion $\boldsymbol{\xi}^*$ which minimizes the bottom weighted sum of squared photometric errors can be obtained by:

$$\boldsymbol{\xi}_i^* = \arg\min_{\boldsymbol{\xi}_i} \sum_{i=1}^{m} \sum_{j=1}^{n} W(r_{ij}) r_{ij}^2(\boldsymbol{\xi}_i) \tag{2}$$

where n is the number of pixels in each image and $W(\cdot)$ is the weighting function determined by Huber weight. To solve the above non-linear weighted optimization problem, the Levenberg-Marquardt algorithm is employed.[9, 10] We represent the Lie algebra parameter as $\boldsymbol{\xi} = [\nu^T \omega^T]^T \in \Re^{6\times1}$ in the tangent space of the Lie group SE(3), where ν and ω are infinitesimal translation and rotation, respectively. A rigid body transformation $T \in SE(3)$ can be defined as a form of exponential map:

$$T(\boldsymbol{\xi}) = \exp\left(\hat{\boldsymbol{\xi}}\right) \tag{3}$$

where $\hat{\xi} \in \Re^{4\times4}$ is a twist matrix of the Lie algebra, $\boldsymbol{\xi}$. Figure 2 shows the process of optimization minimizing the photometric error based on the depth calculation.

In addition, for the computation efficiency, prior to estimating the optimal relative motion with a full-resolution image, we first calculate rough relative motion with the downscaled image, depth information, and an intrinsic matrix of downscaled image. In the proposed algorithm, the optimization is employed for 1/8, 1/4, 1/2 downscaled, and full-resolution image, iteratively. In order to optimize $\boldsymbol{\xi}$, we calculate Jacobian of residual (Eq. (1)) with respect to $\boldsymbol{\xi}$ in analytic approaches. The Jacobian matrix can be written as:[11]

$$J_i^T = \begin{pmatrix} f_x \frac{1}{z'} & 0 \\ 0 & f_y \frac{1}{z'} \\ -f_x \frac{x'}{z'^2} & -f_y \frac{y'}{z'^2} \\ -f_x \frac{x'y'}{z'^2} & -f_y\left(1 + \frac{y'^2}{z'^2}\right) \\ f_x\left(1 + \frac{x'^2}{z'^2}\right) & f_y \frac{x'y'}{z'^2} \\ -f_x \frac{y'}{z'} & f_y \frac{x'}{z'} \end{pmatrix} \begin{pmatrix} \nabla I_{i,x} \\ \nabla I_{i,y} \end{pmatrix} \tag{4}$$

Table 1. **Evaluation results with the EuRoC MAV dataset**

Environment	Relative Pose Error [m/s]			Absolute Trajectory Error [m]			Final Drift Error [%]			Travel Distance [m]	# of keyframe
	Proposed	w/o keyframe	libviso2	Proposed	w/o keyframe	libviso2	Proposed	w/o keyframe	libviso2		
Vicon Room 1 01	0.108	**0.105**	0.560	**0.910**	1.07	4.51	**2.79**	3.79	17.6	37.639	1898
Vicon Room 2 01	**0.169**	0.184	0.314	**2.22**	2.31	5.75	**14.3**	15.1	21.7	32.923	1886
Vicon Room 2 02	0.363	**0.344**	0.924	**2.13**	2.18	3.58	**5.38**	5.80	8.71	74.076	1932

where $\mathbf{x}' = (x', y', z')$ are the transformed 3D coordinates of the corresponding pixel, $\mathbf{x}$. The Jacobian is recalculated at every iteration, because the warped point $\mathbf{x}'$ is reconstructed from $\boldsymbol{\xi}_i$ which is optimization parameter.

3. Evaluation

The proposed stereo visual odometry is validated in the EuRoC MAV dataset.[12] In order to test the performance of the proposed algorithm in the situation where MAV flies indoors, we use the EuRoC MAV dataset. The datasets are captured from a synchronized stereo camera and an inertial measurement unit, which are attached to the MAV, under the various conditions such as slow/fast motion of the MAV and good/bad texture. It also provides an extra dataset for manual calibration. In the EuRoC MAV dataset, the synchronized stereo image pairs and inertial sensor data are collected at 20Hz by a visual-inertial sensor unit attached to an Asctec Firefly hex-rotor. The 6D ground truth is provided with Vicon motion capture system at a rate of 100Hz.

For the quantitative and qualitative evaluation of the proposed algorithm, we use Relative Pose Error(RPE), Absolute Trajectory Error(ATE) and the final drift error.[13] Especially, final drift error is denoted as the ratio of the total travel distance calculated by ground truth. The performance of the proposed algorithm is compared with the well-known existing stereo visual odometry algorithm. The proposed algorithm is implemented in Matlab and runs on 64bit Windows with Intel Core i7-3770@3.4GHz and 8GB memory.

In evaluation results, the proposed algorithm with or without keyframe selection process and libviso2[14] are compared with one another. They are validated with three EuRoC MAV datasets under the various flying environment settings. In Table 1, evaluation results are measured with three kinds of error metrics as mentioned above. As shown in table, the proposed algorithm has much smaller error than libviso2. Also, the keyframe selection algorithm can reduce drift error in terms of final drift error. The number of stereo image pairs tested is 2,000, but the proposed algorithm uses fewer keyframes as shown in the last column of the Table 1. Figures 3 and 4 shows the trajectory estimation result for each stereo visual odometry algorithm with the Vicon Room 2 01 dataset. The proposed algorithm has a small rate of increase in the ATE value as shown in Fig. 4.

4. Conclusion

In this paper, we propose a depth-based stereo visual odometry algorithm which tracks the camera pose using the intensity error of the pixel directly. By using a direct method, the proposed algorithm can improve the robustness to high frequency oscillation or blurring of visual input, thus making it possible to estimate the camera pose accurately. We calculate depth image from stereo calibration and disparity, so it is easy to construct a 3D representation of pixels. In the research, we use a stereo camera in order to compensate the scale drift error, and a selected keyframe has a role as a reference image to estimate the relative motion of a camera resulting in reduction of drift error when compared with a popular existing algorithm.

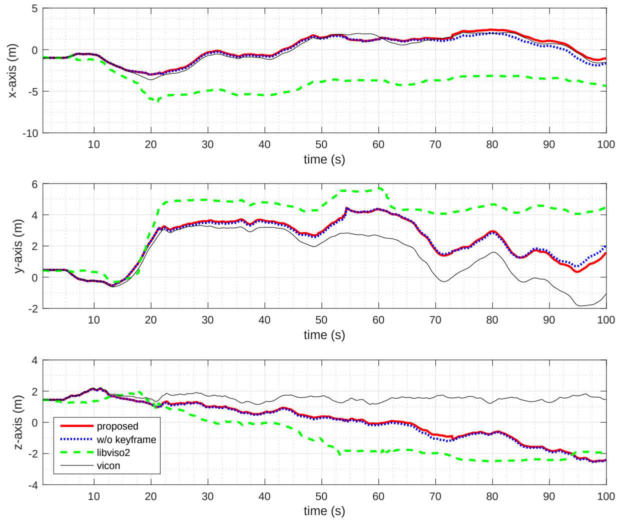

Fig. 3. **The trajectory estimation of each axis for each visual odometry method with Vicon Room 2 01 dataset.**The proposed algorithm with keyframe selection has better performance than the others.

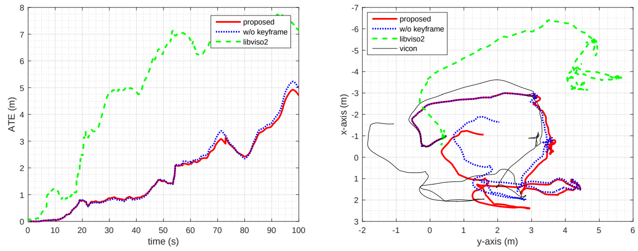

Fig. 4. **The trajectory estimation results for each visual odometry method with Vicon Room 2 01 dataset.** (left) Absolute trajectory error values over time are shown. It shows drift errors of each algorithm. (right) The estimated trajectory with the proposed algorithm with (red) or without (blue) keyframe selection, and libviso2 (green).

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT & Future Planning (MSIP) (No. 2014034854 / 2014M1A3A3A02034854).

This work is financially suppertedby Ministry of Land, Infrastructure and Transport as 「NALDA Project」.

References

1) Nistér, David, Oleg Naroditsky, and James Bergen. ”Visual odometry.” Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Vol. 1. IEEE, 2004.

2) Scaramuzza, Davide, and Friedrich Fraundorfer. ”Visual odometry [tutorial].” IEEE Robotics & Automation Magazine 18.4 (2011): 80-92.

3) Engel, Jakob, Jörg Stückler, and Daniel Cremers. ”Large-scale direct SLAM with stereo cameras.” Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. IEEE, 2015.

4) Klein, Georg, and David Murray. ”Parallel tracking and mapping on a camera phone.” Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on. IEEE, 2009.

5) Davison, Andrew J., et al. ”MonoSLAM: Real-time single camera SLAM.” IEEE transactions on pattern analysis and machine intelligence 29.6 (2007): 1052-1067.

6) Kerl, Christian, Jürgen Sturm, and Daniel Cremers. ”Robust odometry estimation for RGB-D cameras.” Robotics and Automation (ICRA), 2013 IEEE International Conference on. IEEE, 2013.

7) Newcombe, Richard A., Steven J. Lovegrove, and Andrew J. Davison. ”DTAM: Dense tracking and mapping in real-time.” 2011 international conference on computer vision. IEEE, 2011.

8) Hirschmuller, Heiko. ”Accurate and efficient stereo processing by semi-global matching and mutual information.” 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05). Vol. 2. IEEE, 2005.

9) Levenberg, Kenneth. ”A method for the solution of certain non-linear problems in least squares.” Quarterly of applied mathematics 2.2 (1944): 164-168.

10) Marquardt, Donald W. ”An algorithm for least-squares estimation of nonlinear parameters.” Journal of the society for Industrial and Applied Mathematics 11.2 (1963): 431-441.

11) Kerl, Christian. Odometry from rgb-d cameras for autonomous quadrocopters. Diss. Masters thesis, Technical University Munich, Germany, 2012.

12) Burri, Michael, et al. ”The EuRoC micro aerial vehicle datasets.” The International Journal of Robotics Research (2016): 0278364915620033.

13) Sturm, Jürgen, et al. ”A benchmark for the evaluation of RGB-D SLAM systems.” 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012.

14) Geiger, Andreas, Julius Ziegler, and Christoph Stiller. ”Stereoscan: Dense 3d reconstruction in real-time.” Intelligent Vehicles Symposium (IV), 2011 IEEE. IEEE, 2011.