Journal of Imaging Science and Technology® 67(6): 1–6, 2023.

Camera Motion Estimation Method using Depth-Normalized Criterion

Seok Lee
Mechatronics Department, KOREATECH, Cheonan-si, Chungcheongnam-do, South Korea
E-mail: leeseok@koreatech.ac.kr

Abstract. *For translationally moving objects with fixed cameras, such as robots and cars, blurring can often be more pronounced in objects that are closer to the camera. A depth-normalized, least-squares objective function is proposed for the simultaneous recovery of shape and motion parameters from optical flow, together with an efficient iterative optimization algorithm. Simulation and experiments demonstrate that for scenes with sufficient depth variation, our algorithm provides robust, statistically consistent estimates of shape and motion.* © *2023 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2023.67.6.060403]

1. INTRODUCTION

Motion and depth estimation using vision sensors, otherwise known classically as the structure from motion problem (SFM), is an essential component of mobile robot localization, mapmaking, and navigation. While the SFM problem is well-understood and generally considered to be solved, mobile robots are often equipped with off-the-shelf low performance sensors – particularly with the proliferation of low-cost mobile robots for the mass market – and operate in unstructured environments under uneven lighting conditions which makes the problem to be solved based on statistical principles such as Kalman filter or Markov method [1, 2], and this makes the estimation problem still relevant and challenging [3–7].

In the classical structure from motion (SFM) literature, it is now well recognized that noise in the image velocities, together with the presence of just a few outliers, can significantly degrade the estimates of depth and motion. Fermuller et al. [8] and Simoncelli et al. [9] have investigated the probabilistic and statistical characteristics of optical flow measurements, while Daniilidis and Spetsakis [10] offer a comprehensive framework addressing various sources of error in motion estimation (e.g., statistical bias, correlated noise, geometric instabilities).

One factor contributing to this noise sensitivity is that most existing SFM algorithms treat the entire set of optical flow measurements uniformly, regardless of the distance from the camera, or whether the translation or rotation component of the motion is more dominant. In typical video scenes taken in urban settings, for example, it is quite common for objects to be moving at a wide range of camera depths. In the case of translational motions, the magnitude of optical flow is inversely proportional to depth, and blurring which is caused by camera exposure is often more pronounced for objects that are closer to the camera, while the optical flow measurements of extremely distant points tend to be dominated by noise. It would thus seem reasonable to rely more on optical flow measurements that are sufficiently distant from the camera to minimize the effects of blurring, while ensuring an appropriate signal-to-noise ratio. A closely related idea is that of Heeger [11], who formulated image flow uncertainty in such a way that it increases with flow magnitude.

This paper presents a depth-normalized criterion for simultaneously recovering velocity and depth information from optical flow data, together with an efficient iterative algorithm for its optimization. Intended for scenarios where near points are subject to greater blurring, our objective function normalizes the data such that the optical flow measurements from distant points are given proportionally greater weight. We present an efficient cyclic coordinate descent algorithm for obtaining the shape and motion estimates. Finally, extensive simulation and experimental studies are conducted to assess the performance of our algorithm, and results show that, for scenes with sufficient depth variation, our algorithm leads to more robust and accurate shape and motion estimators.

2. PROBLEM FORMULATION

2.1 *Camera Model & Measurements*

We assume a standard perspective projection model for a camera with unit focal length. The image velocity of the camera motion in this case becomes

$$u(p) = \lambda(p)A(p)v + B(p)\omega + n(p), \tag{1}$$

where $u(p) = (u_x(p), u_y(p))^T$ is the two-dimensional image velocity vector at image position $p = (p_x, p_y, 1)^T$, $v = (v_x, v_y, v_z)^T$ is the camera's translational velocity, $\omega = (\omega_x, \omega_y, \omega_z)^T$ is its angular velocity, and the scalar $\lambda(p)$ is the inverse scene depth at image point p. The term $n(p) = (n_x(p), n_y(p))^T$ denotes noise, and

$$A(p) = \begin{bmatrix} 1 & 0 & -p_x \\ 0 & 1 & -p_y \end{bmatrix}, \tag{2}$$

Received June 28, 2023; accepted for publication Nov. 11, 2023; published online Dec. 12, 2023. Associate Editor: Steven Simske.
1062-3701/2023/67(6)/060403/6/$25.00

$$B(p) = \begin{bmatrix} -p_x p_y & 1+p_x^2 & -p_y \\ -(1+p_y^2) & p_x p_y & p_x \end{bmatrix}. \quad (3)$$

Given a collection of n optical flow measurements $\{(p_1, u_1), \ldots, (p_n, u_n)\}$, the objective is to estimate the translational and angular velocities v and ω, and the inverse depths $\lambda_1, \ldots, \lambda_n$ associated with each of the image points $p_1, \ldots, p_n$ in some optimal fashion. It is well known that since $\lambda(p)$ and v appear as a product in Eq. (1), it is not possible to determine their magnitudes; we therefore adopt the standard practice of assuming $\|v\| = 1$.

Zhang and Tomasi [11] have shown that nonisotropic noise models for optical flow can lead to statistically inconsistent motion parameter estimates, in the sense of infinite-sample unbiasedness and finite-sample convergence - intuitively, the estimates fail to improve in accuracy with more optical flow measurements. Their study also highlights the sometimes fatal consequences caused by inappropriate transformations of the original SFM problem formulation, particularly those based on epipolar geometry. Epipolar methods have the advantage of decoupling the depth and motion estimation problems; by algebraically eliminating depth from the objective function via the epipolar constraint, the dimension of the ensuing optimization problem is significantly reduced. The depth parameters can moreover be recovered by a simple postprocessing procedure involving a singular value decomposition. One study [12] emphasized that the motion-depth decoupling achieved in the various epipolar methods are due to transformations of the fundamental SFM problem and also cited several examples of popular epipolar geometry-based SFM estimators that fail to be statistically consistent *e.g.*,[13–15].

Zhang and Tomasi [12] further show that under the assumption that the errors of the optical flow measurements are independent, identically distributed, and isotropic (in the sense of being rotationally symmetric), the estimator given by

$$\text{argmin}_{\omega,v} \sum_{i=1} \inf_{\lambda_i} \|A_i([\omega]\, p_i + \lambda_i v) - u_i\|^q, \quad (4)$$

where $\omega \in \Re^3$, $v \in S^3$ ($\|v\| = 1$), $q \geq 1$ and $\|\cdot\|$ denotes the Euclidean two-norm, is statistically consistent (Their objective function is presented in slightly more general form than the one given here). An efficient iterative Gauss-Newton algorithm is also derived.

2.2 *Depth Normalized Objective Function*

Figure 1 illustrates the image blurring that can occur in typical dynamic urban scenes; this image was taken from a moving car at 1/50 shutter speed. The direction of camera movement is perpendicular to optical axis, and magnitude of motion field is inversely proportional to depth of object scene. During camera exposure, the image is blurred by the motion field which is induced by camera translation. Images captured from sensor always contain this motion blur because camera exposure time is finite and larger than zero.

Figure 1. Blurring and optical flow noise with respect to camera depth value.

In the case of mobile robot which moves in linear translation, the accuracy of visual navigation is affected by this motion blur in each frame because it uses consecutive camera frames to estimate motion and depth information.

To compensate for this particular type of blurring phenomena, we propose a modified version of the objective function (4) that weights the optical flow measurements according to depth:

$$J(\omega, v, \lambda) = \sum_{i=1}^{n} \frac{1}{\lambda_i^2} \|A_i([\omega]\, p_i + \lambda_i v) - u_i\|^2. \quad (5)$$

Informally, the inverse depth scaling has the effect of "undoing" the perspective projection before considering the noise. To ensure that the flow measurements are of sufficient signal-to-noise ratio, in practical implementations, one would discard measurements that are beyond a certain threshold depth; these and other practical issues are discussed in detail later.

3. SOLUTION

As is common in the SFM literature, we focus on the case $v \neq 0$ because the $v = 0$ case can be easily detected and treated separately. The optimal λ can be determined parametrically as a function of ω and v from the first-order necessary conditions for optimality, *i.e.*, by setting gradient equals to zero. This leads to

$$\lambda_k = \frac{\|u_k - B(p_k)\omega\|^2}{(u_k - B(p_k)\omega)^T A(p_k) v} \quad (6)$$

Given values for ω and v, the λ that minimizes the cost function (5) is given by the above. By substituting $\lambda(\omega, v)$ above back into (5), the cost function becomes, after some manipulation,

$$J(\omega, v) = \sum_{i=1}^{n} \left\| \left(I - \frac{(u_i - B(p_i)\omega)(u_i - B(p_i)\omega)^T}{\|u_i - B(p_i)\omega\|^2} \right) A(p_i) v \right\|^2. \quad (7)$$

We use the following notation:

$$Q_i(\omega) = A(p_i) - \frac{(u_i - B(p_i)\omega)(u_i - B(p_i)\omega)^T}{\|u_i - B(p_i)\omega\|^2} A(p_i),$$

$$Q(\omega) = \begin{bmatrix} Q_1(\omega) \\ \vdots \\ Q_n(\omega) \end{bmatrix}.$$

Note that $Q_i(\omega) \in \Re^{3\times 3}$ and $Q(\omega) \in \Re^{2n\times 3}$. The objective function can now be written as

$$J(\omega, v) = \|Q(\omega)v\|^2. \tag{8}$$

If ω is given, using a Lagrange multiplier argument, one can show that the optimal v is given by the unit-length eigenvector of $Q^T Q$ corresponding to the smallest eigenvalue where the cost function is symmetric with respect to v, in the sense that both v and $-v$ lead to identical values of the objective function.

Instead of attempting to simultaneously minimize the cost function with respect to ω and v, we minimize sequentially over the two parameters as follows:

- Let $k = 0$ and choose any initial value $\omega_k \in \Re^3$;
- Iterate the following:
  - v_k = unit-length eigenvector of $Q^T Q$ corresponding to the minimal eigenvalue;
  - $\omega_{k+1} = \underset{\omega\in\Re^3}{\text{argmin}} J(\omega, v_k)$, where v_k is obtained from the previous step;
  - $k = k+1$.

Under various compactness and uniqueness assumptions one can show via the global convergence theorem (see [16, 17]) that the above cyclic coordinate descent (CCD) algorithm is assured of converging to meaningful local minima. We do not address the details here but refer the reader to [18, 19] and the previous references for applications of the global convergence theorem in vision settings, and a discussion of the subtleties.

We now examine in more detail the conditional problem of minimizing $J(\omega, v)$ given $v \in S^2$; we denote this conditional objective function by $J(\omega|v)$. Defining

$$b_i(\omega) = u_i - B(p_i)\omega, \tag{9}$$

$J(\omega|v)$ can be written after some manipulation as

$$J(\omega|v) = \sum_{i=1}^{n} \left(\|A(p_i)v\|^2 - \frac{b_i^T A(p_i) v v^T A^T(p_i) b_i}{\|b_i\|^2} \right).$$

Ignoring the $\|A(p_i)v\|^2$ term (since v is assumed given), and defining

$$R_i(v) = A(p_i) v v^T A^T(p_i), \tag{10}$$

we have the following sum-of-ratios quadratic fractional programming problem:

$$\min_{\omega\in\Re^3} J(\omega|v) = -\sum_{i=1}^{n} \frac{b_i^T R_i b_i}{b_i^T b_i}. \tag{11}$$

Each R_i is symmetric, positive semidefinite, and of rank one. The analytic gradient of $J(\omega|v)$ is useful for numerical optimization purposes:

$$\frac{\partial J(\omega|v)}{\partial \omega} = \sum_{i=1}^{n} \left(\frac{u_i^T R_i B(p_i) - \omega^T B^T(p_i) B(p_i)}{\|b_i\|^2} - \frac{b_i^T R_i b_i}{\|b_i\|^4} (u_i^T B(p_i) - \omega B^T(p_i) B(p_i)) \right). \tag{12}$$

With this gradient, any number of standard optimization algorithms and specialized algorithms for fractional programming are at our disposal [20].

4. EXPERIMENTAL RESULTS

4.1 *Synthetic Data*

Experiments with synthetic data have been performed with our proposed algorithm, and the results are compared with the algorithms of Zhang and Tomasi [12] and Soatto and Brockett [21]; the latter developed a cyclic descent optimization algorithm for a standard epipolar geometry-based motion estimation criterion. 50 feature points are randomly generated from a uniform distribution in a three-dimensional $120 \times 120 \times 120$ region. These points are assumed to belong to a single rigid body moving with translational velocity (1, 3, 2) and angular velocity (−1, 0.5, 1.5). Corresponding optical flow measurements are obtained via perspective projection. Independent uncorrelated Gaussian noise is then added to the measurements after scaling the noise by depth. In our simulations noise levels are successively increased up to 50% of the average optical flow magnitudes.

Spherical velocity errors are measured according to

$$d(v_{act}, v_{est}) = \cos^{-1}(v_{act}, v_{est}), \tag{13}$$

where $v_{act} \in S^2$ denotes the actual velocity, and $v_{est} \in S^2$ denotes the estimated value obtained from the optimization. Physically, this metric corresponds to the angle between v_{act} and v_{est}; that this definition satisfies the distance metric axioms can be straightforwardly verified. Linear velocity errors are measured in the standard way in terms of the Euclidean two-norm. In the optimization procedure we use the stopping criterion

$$\frac{|J_{k+1}(\omega, v) - J_k(\omega, v)|}{|J_k(\omega, v)|} < \epsilon, \tag{14}$$

where ϵ is on the order of 10^{-6}.

We first examined whether increasing the number of feature points increases the accuracy of the linear and angular velocity estimates produced by our depth-normalized

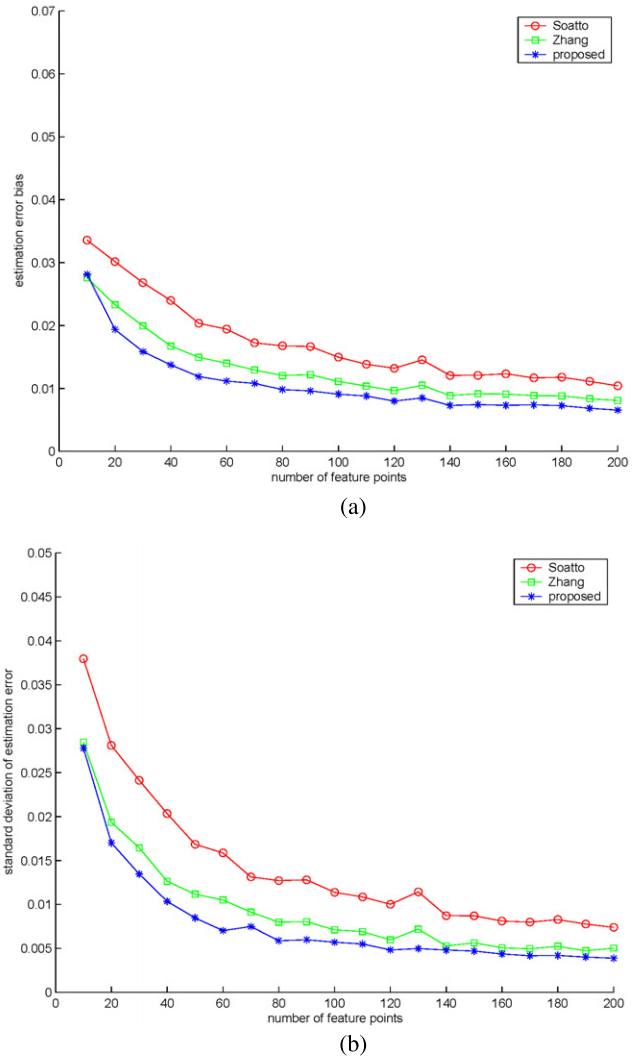

Figure 2. (a) Estimation error bias and (b) standard deviation versus number of feature points.

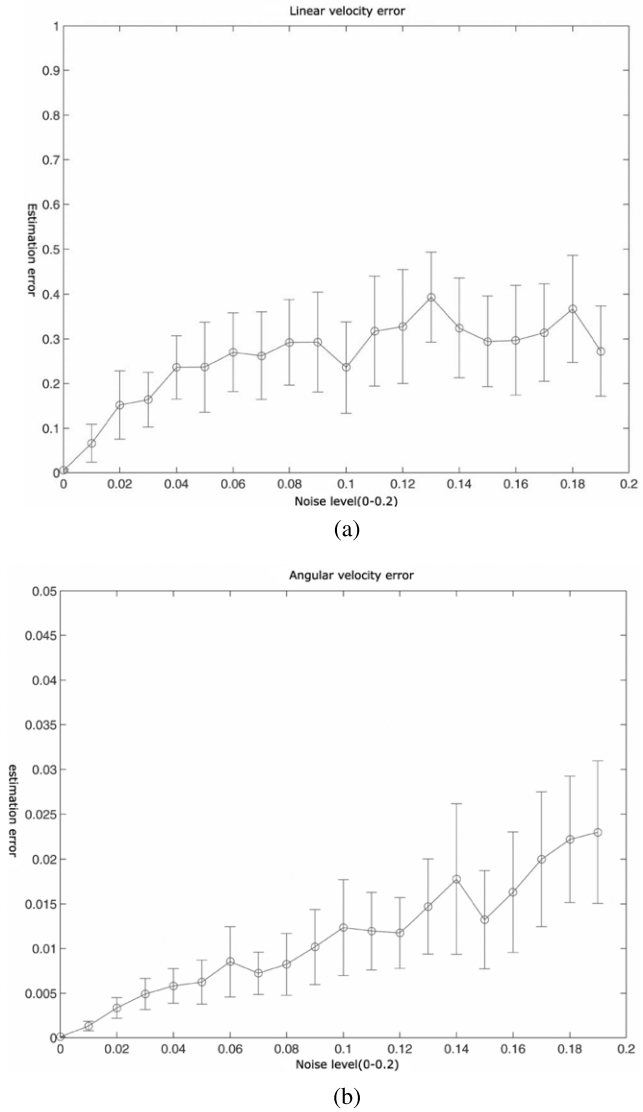

Figure 3. (a) Linear and (b) angular velocity errors as a function of noise level.

criterion. Adding depth-scaled zero-mean Gaussian noise with 0.1 standard deviation to the optical flow measurements, we examined both the error and standard deviation of the linear and angular velocity estimates as a function of the number of feature points. The feature points are increased from 1,000 to 10,000 in increments of 1000. Figure 2 shows the results of our algorithm for synthetic data, averaged over 50 sample trials; the top graph illustrates the estimation error bias, while the bottom shows the standard deviation, both as a function of the number of feature points. Our results are also compared with those obtained using the Zhang-Tomasi ($Z-T$) and Soatto-Brockett ($S-B$) algorithms. All three algorithms display a distinct trend of decreasing bias and variance with increasing number of feature points; however, our depth-normalized criterion shows the most rapid decrease.

We then examined the noise sensitivity of our depth-normalized algorithm. 30 feature points are used, and noise levels are successively increased up to 50% of the average optical flow magnitudes (corresponding to absolute values of around 0.2). Figure 3 illustrates the linear and angular velocity estimation errors and standard deviation as a function of increasing noise levels. The errors are again obtained as the average of 50 trials, with the ranges indicating plus-minus one standard deviation. The errors can be seen to increase in approximately linear fashion as noise levels are increased.

Table I. Computation times for the three algorithms.

	Proposed	$Z-T$	$S-B$
Time (s)	0.73	0.70	0.49
Iterations	7.2	5.8	4.8

Table I lists the average computation times for the three algorithms. All the algorithms were implemented in

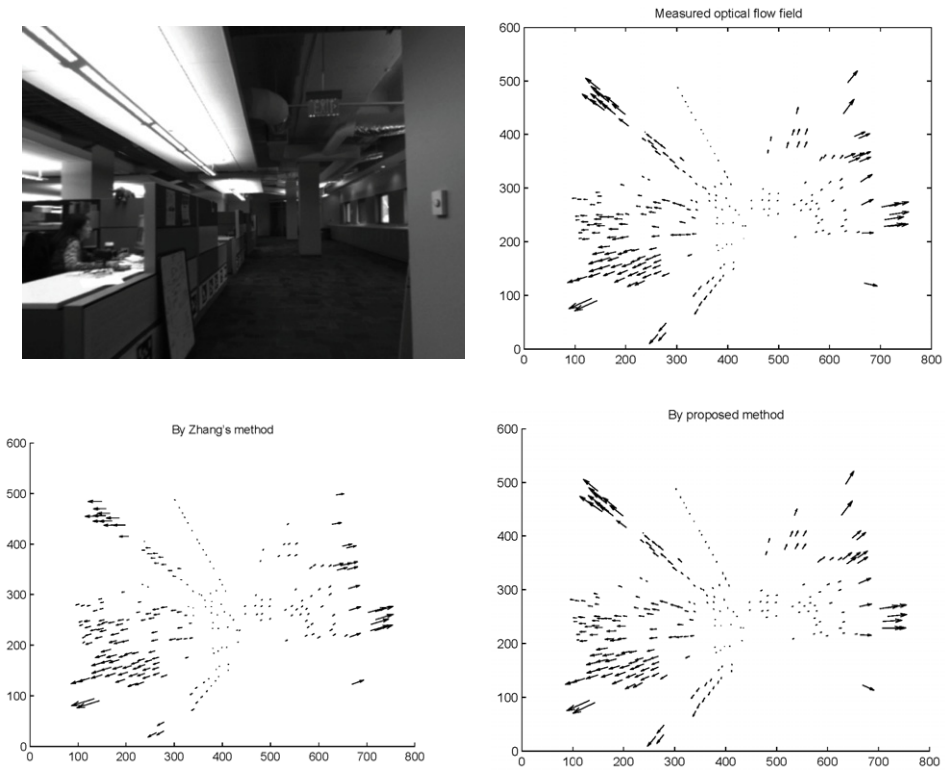

Figure 4. Experimental results for scene 1.

the Microsoft Visual C++. Feature point detection and optical flow calculation were performed using the appropriate OpenCV routines, and the nonlinear optimization routines from IMSL's PC version were used for numerical optimization, in conjunction with an internally developed matrix computation library (RMatrix).

Not surprisingly, our proposed algorithm was the slowest, followed closely by the Z–T algorithm. Since the above two algorithms explicitly solve for the depth parameters in the optimization, this result is not unexpected. The Z–T algorithm, which eliminates the depth parameters altogether via the epipolar constraint, was the fastest of the three algorithms.

4.2 *Real Images*

We then evaluated our algorithm on a series of scenes that are captured from a Point Grey Flea camera mounted on a Pioneer PeopleBot; this camera was used to vary the exposure time and iris so as to produce a range of blurring effects. The scene depicted in Figures 4, 5 contains objects at depths of up to 30 m. We deliberately obtained motion sequences at slow shutter speeds to capture the blurring effect and raise noise levels. Optical flow measurements were obtained using the OpenCV pyramidal implementation of the iterative Lucas-Kanade method; the feature points were also extracted using the OpenCV library.

The camera underwent a linear translation directly toward the objects along the robot and the optical flow measurements are shown in the upper-right figure. One can observe the relatively large number of incorrect optical flow vectors for the proximal object; the errors for the proximal object are more pronounced than for distal objects. Comparing the optical flow fields estimated using our proposed algorithm with that obtained from the Z–T algorithm, our algorithm shows better performance; the directional errors present in the measured flow field are largely corrected using our depth-normalized criterion.

5. CONCLUSION

One factor contributing to the noise sensitivity of existing SFM algorithms is that the optical flow measurements of all points, regardless of their depth, are treated with the same degree of fidelity. This paper has proposed a depth-normalized criterion that places a greater weight on the optical flow measurements at increased depths. The underlying premise is that for mobile robots and cars with fixed cameras that are traveling linearly in typical scenes, particularly in urban environments, blurring (and thus more noise) is often more pronounced in objects that are closer to the camera. We derived an efficient cyclic optimization algorithm for estimating the velocity and depth parameters. Experiments with both synthetic data and real images suggest that for scenes with sufficient depth variation, and in which translational motions are dominant, our depth-normalized criterion leads to improved estimates of the velocity and depth. The proposed motion estimation method is not superior in terms of computational efficiency because the depth value is obtained during the optimization process, while other reference algorithms calculate it explicitly.

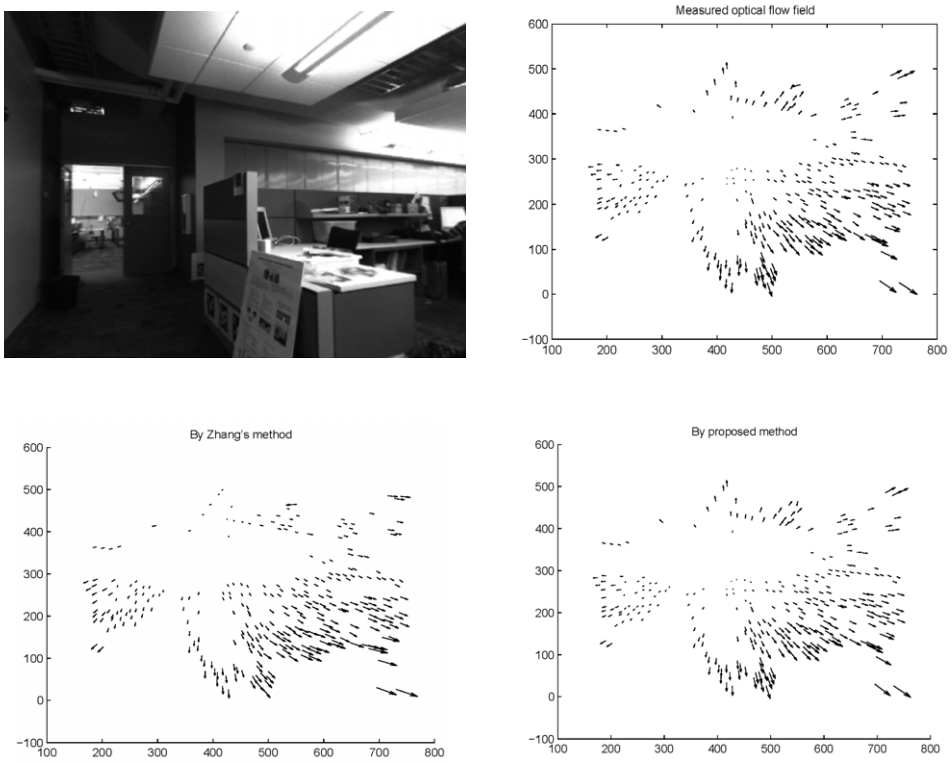

Figure 5. Experimental results for scene 2.

Future work should explore more computationally efficient motion estimation algorithms using the proposed depth normalization criterion. We are also working to improve the current implementation of the proposed method, including removing dependencies on internally developed libraries.

ACKNOWLEDGMENT

This paper was supported by Education and Research promotion program of KOREATECH in 2022.

REFERENCES

[1] M. Bashiri, H. Vatankhah, and S. S. Ghidary, "Hybrid adaptive differential evolution for mobile robot localization," J. Intel. Serv. Robotics **5**, 99–107 (2011).
[2] D. R. Parhi and S. Kundu, "Navigation control of underwater robot using dynamic differential evolution approach," Proc IMechE Part M: J Engineering for the Maritime Environment **231**, 284–301 (2017).
[3] J. Kim, C. Park, and I. S. Kweon, "Vision-based navigation with efficient scene recognition," J. Intel. Serv. Robotics **4**, 191–202 (2011).
[4] L. Niu, S. Smirnov, J. Mattila, A. Gotchev, and E. Ruiz, "Robust pose estimation with a stereoscopic camera in harsh environments," *Proc. IS&T Electronic Imaging: Intelligent Robotics and Industrial Applications using Computer Vision 2018* (IS&T, Springfield, 2018), pp. 126-1–126-6.
[5] M. B. Alatise and G. P. Hancke, "A review on challenges of autonomous mobile robot and sensor fusion methods," IEEE Access **8**, 39830–39846 (2020).
[6] X. Zhang, W. Wang, X. Qi, Z. Liao, and R. Wei, "Point-plane SLAM using supposed planes for indoor environments," Sensors **19**, 3795 (2019).
[7] J. A. Placed and J. A. Castellanos, "A deep reinforcement learning approach for active SLAM," Appl. Sci. **10**, 8386 (2020).
[8] C. Fermuller, D. Shulman, and R. Pless, "The statistics of optical flow," Comput. Vis. Image Underst. **82**, 1–32 (2001).
[9] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger, "Probability distributions of optical flow," *IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 1991).
[10] K. Daniilidis and M. Spetsakis, "Understanding noise sensitivity in structure from motion," in *Visual Navigation*, edited by Y. Aloimonos (Psychology Press, East Sussex, 1996), pp. 61–88.
[11] D. J. Heeger, "Optical flow using spatiotemporal filters," Int. J. Comput. Vis. **1**, 279–302 (1988).
[12] T. Zhang and C. Tomasi, "On the consistency of instantaneous rigid motion estimation," Int. J. Comput. Vis. **46**, 51–79 (2002).
[13] A. Bruss and B. Horn, "Passive navigation," Comput. Graph. Image Process. **21**, 3–20 (1983).
[14] X. Zhuang, T. Huang, N. Ahuja, and R. Haralick, "A simplified linear optic flow-motion algorithm," Comput. Vis. Graph. Image Process. **42**, 334–344 (1988).
[15] A. D. Jepson and D. J. Heeger, "Linear subspace methods for recovering translation direction," in *Spatial Vision in Humans and Robots*, edited by L. Harris and M. Jenkins (Cambridge University Press, Cambridge, 1993), pp. 39–62.
[16] W. Zangwill, *Nonlinear Programming: A Unified Approach* (Prentice-Hall, Englewood Cliffs, 1969).
[17] D. G. Luenberger, *Linear and Nonlinear Programming* (Addison Wesley, Boston, 1989).
[18] S. Mahamud, M. Hebert, Y. Omori, K. McHenry, and J. Ponce, "Provably-convergent iterative methods for projective structure from motion," *IEEE Int'l. Conf. Computer Vision & Pattern Recognition* (IEEE, Piscataway, NJ, 2001), pp. 1018–1025.
[19] S. Gwak, J. Kim, and F. C. Park, "Numerical optimization on the Euclidean group with applications to camera calibration," IEEE Trans. Robot. Autom. **19**, 65–74 (2003).
[20] S. Schaible and J. Shi, "Fractional programming: the sum-of-ratios case," Optimization Methods Softw. **18**, 219–229 (2003).
[21] S. Soatto and R. Brockett, "Optimal structure from motion: Local ambiguities and global estimates," Int. J. Comput. Vis. **39**, 195–228 (2000).