Adaptive Essential Matrix Based Stereo Visual Odometry with Joint Forward-Backward Translation Estimation

Huu Hung Nguyen[1(✉)], Quang Thi Nguyen[1], Cong Manh Tran[1], and Dong-Seong Kim[2]

[1] Le Quy Don Technical University, 236 Hoang Quoc Viet, Hanoi, Vietnam
{hungnh.isi,thinq.isi}@lqdtu.edu.vn, manhtc@gmail.com
[2] Kumoh National Institute of Technology, Gumi-si, South Korea
dskim@kumoh.ac.kr

Abstract. Visual Odometry is widely used for recovering the trajectory of a vehicle in an autonomous navigation system. In this paper, we present an adaptive stereo visual odometry that separately estimates the rotation and translation. The basic framework of VISO2 is used here for feature extraction and matching due to its feature repeatability and real-time speed on standard CPU. The rotation is accurately obtained from the essential matrix of every two consecutive frames in order to avoid the affection of the stereo calibration uncertainty. With the estimated rotation, translation is rapidly calculated and refined by our proposed linear system with non-iterative refinement without the requirement of any ground truth data. The further improvement of the translation by joint backward and forward estimation is also presented in the same framework of the proposed linear system. The experimental results evaluated on the KITTI dataset demonstrate around 30% accuracy enhancement of the proposed scheme over the traditional visual odometry pipeline without much increase in the system overload.

Keywords: Visual odometry · Essential matrix · Non-iterative translation estimation · Forward-backward translation estimation

1 Introduction

Visual Odometry(VO) [1] is one of the important parts in robotics research, especially for autonomous navigation. With the unavailability of GPS signals such as indoor extra-terrestrial and in space, image-based localization becomes necessary. Specifically, a single movement between previous and current images is estimated by resolving geometric constraints. Subsequently, the full camera trajectory is finally recovered via the accumulation of these movements. Recently, the

N.-S. Vo and V.-P. Hoang (Eds.): INISCOM 2020, LNICST 334, pp. 127–137, 2020.
https://doi.org/10.1007/978-3-030-63083-6_10

survey [2] classified VO in different approaches such as monocular/stereo camera-based, geometric/learning-based and feature/appearance-based. The feature-based VO pipeline has a long history and has been detailed in Nister's [1] work. Scaramuzza and Fraundorfer conducted a comprehensive review of feature-based VO [3,4]. Accordingly, relative ego-motion between the two frames was obtained by three following major approaches

- **2D-to-2D**: Motion is estimated only from 2D feature correspondences.
- **3D-to-3D**: Motion is estimated only from 3D feature correspondences.
- **3D-to-2D**: Motion is estimated from 3D features in one frame and their corresponding 2D features in other.

The first approach, 2D-to-2D,is a methodology that recovers the rotation and translation direction from the essential matrix computed from 2D feature correspondences using the epipolar constraint. Nister proposed an efficient implementation [5] for the minimal case solution with five 2D correspondences and that has become the standard for 2D-to-2D motion estimation due to the efficiency in the presence of outliers. The second approach, 3D-to-3D, computes the camera motion by determining the aligning transformation of the two 3D feature sets that minimizes the Euclidean distance between the two sets of 3D features. As shown in [6], the minimal case solution involves three 3D-to-3D non-collinear correspondences, which can be used for robust estimation in the presence of outliers. 3D-to-2D method is well-known as perspective from n points (PnP). The pose is obtained via iteratively minimizing the summation of projection error between the projected points of 3D features and corresponding 2D observations. The minimal case involving 3D-to-2D correspondences in [7] is called perspective from three points ($P3P$). With $n \geqslant 6$, there is a simple and straightforward solution for PnP by solving the direct linear transformation (DLT) [8]. The conventional framework VISO2 [9] applied PnP approach with a robust against outliers. They adopted PnP using 3 randomly drawn correspondences into a RANSAC scheme, by first estimating (R, t) for 50 times independently. All inliers of the winning iteration were then used for refining the parameters, yielding the final transformation (R, t).

Since VO works by estimating the camera path incrementally (pose after pose), then over time, the errors introduced by each frame-to-frame motion accumulate. This generates a drift of the estimated trajectory from the real path. For some applications, it is utmost important to keep drift as small as possible, which can be done through local optimization over the last m camera poses. This approach called sliding window bundle adjustment or windowed bundle adjustment has been used in several works. However, it takes additional computational time because of being an iterative method. So it only is used as the final step for refining or executed at some special location. For real-time applications, reducing computational cost is important so proposing the fast and accurate frame to frame VO is still an active research area.

Note that, 3D-to-3D and 3D-to-2D approaches require triangulation of 3D points from 2-D image correspondences which are determined by intersecting back-projected rays from 2D image correspondences of at least two image frames.

In perfect conditions, these rays would intersect in a single 3D point. However, they never intersect because of image noise, camera model and calibration errors as well as feature matching uncertainty. Therefore, the point at a minimal distance in the least-squares sense from all intersecting rays can be taken as an estimate of the 3D point position. As pointed out by Nister [1], the 2D-to-2D method and 3D-to-2D method are evaluated to be more accurate than 3D-to-3D methods because 3D-to-3D minimizes feature position error whereas 3D-to-2D approach minimizes re-projection error or 2D-to-2D approach minimizes the Sampson error. In the case of using the 2D-to-2D approach, we do not need triangulation to calculate the rotation and scalable translation. However, we need to use triangulation for computing absolute translation. There is an easy way to obtain scale from relative distances between any combination of two 3D points or exploiting the trifocal constraint between 3 view matches of 2D features or iteratively minimizing re-projection error with known rotation for features on pairs of stereo images.

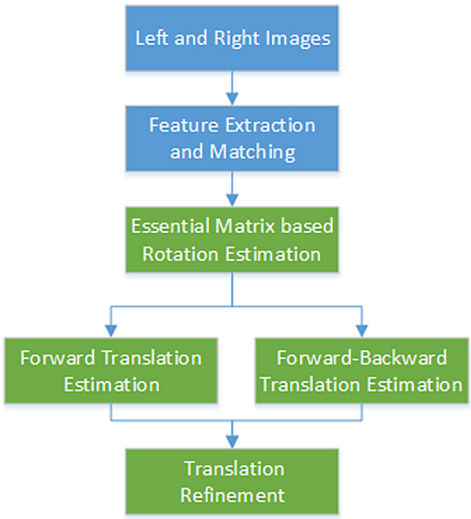

Fig. 1. Our proposed algorithm. Utilizing the feature extraction and matching component of VISO2 library and replacing the pose estimation part based on PnP by our proposed adaptive stereo VO with essential matrix based rotation estimation and join forward-backward translation estimation

In this paper, we propose adaptive stereo camera which estimates the transformation of two consecutive frames separately. The proposed algorithm is depicted in Fig. 1. Here, rotation is extracted from essential matrix estimated by using five point algorithm with preemptive RANSAC and translation is computed by solving a novel linear equation system modified from the re-projection equation. The proposed approach has following benefits:

- Accurate rotation from essential matrix estimation that avoids uncertainty of stereo calibration.
- Translation is estimated directly without iterative optimization for both initiation and final refinement joint forward and backward approaches.

The paper is organized as follows. Section 2 briefly provides a description of feature extraction and matching. Section 3 describes the proposed visual odometry approach based on the essential matrix estimation and non-iterative translation calculation. Section 4 presents the results of KITTI dataset in comparison to VISO2. Section 5 provides the concluding marks.

2 Feature Extraction

The input of our visual odometry is feature correspondences between four images in the previous and current stereo camera frames. We took advantage of feature detector and matching (active matching) proposed by Chli [9] because of its feature repeatability and speed. It was employed as open-source by Geiger in the VISO2 library [10]. In particular, firstly, 5×5 blob and corner masks were used to filter the input image to extract four feature classes: blob min/max, corner min /max. Additionally, feature matching was done by comparing the 11×11 block windows of horizontal and vertical Sobel filter responses to give two features using the sum of absolute differences (SAD) error metric. To speed-up, the matching process, sum over a sparse set of 16 locations was used instead of being summed over the whole block window. Note that, the matching process was done on the same class (blob max/min, corner max/min) to reduce the computational cost without reducing feature matching quality. Some of the outliers were rejected by circular matching [11], suggesting that each feature needs to be matched between left and right images of two consecutive frames, requiring four matches per feature. Finally, the bucketing technique is used to divide the corresponding features into 50×50 grids and selecting only a limited number of features in each bucket. This step guarantees the uniform distribution of selected features along the z-axis, the roll axis of the vehicle. It means that both close and far features are used for pose estimation resulting in high accuracy of vehicle trajectory. This feature detector and matching have been used for several visual odometry approaches such as [12] and [13] that achieved good performance on the KITTI dataset.

3 Proposed Pose Estimation

Single movement between previous and current frames is separated into two parts. Firstly, the five-point algorithm is performed to obtain the essential matrix and then rotation is extracted. Secondly, the translation initially is estimated by one point RANSAC by close features and finally is obtained by forward-backward refinement. Different from conventional approaches using iterative optimization such as [13], we proposed a closed-form linear equation for both initial and final estimation.

3.1 Rotation Estimation

The geometric relation between two consecutive frames of a calibrated camera is described by the so-called essential matrix E which contains the camera motion parameters up to a unknown scale factor for translation. It is represented as following form:

$$E = T^{\times} R \tag{1}$$

where skew matrix $T^{\times}$ is rewritten in detail as follow

$$T^{\times} = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \tag{2}$$

Each correspondence of two image frames satisfies the epipolar constraint

$$p^T E q = 0 \tag{3}$$

Where a 2D feature, p in previous frame corresponds to another 2D feature, q, in current frame. Essential matrix E has two additional properties

$$det(E) = 0 \tag{4}$$

And

$$2EE^T E - tr(EE^T)E = 0 \tag{5}$$

Note that, essential matrix E is 3×3 matrix with 8 unknown variables with an un-observable scale can be solved by five point equations by search solution of root of tenth degree polynomial proposed by Nister [5]. Each of the five point correspondences gives rise to a constraint (3). It can be also rewritten in a linear equation formular as follows

$$\hat{a}\hat{E} = 0 \tag{6}$$

where

$$\hat{a} = [p_1 q_1, q_1 q_2, p_1, p_2 q_1, p_2 q_2, p_2, q_1, q_2, 1] \tag{7}$$

and

$$\hat{E} = [E_{11}, E_{12}, E_{13}, E_{21}, E_{22}, E_{23}, E_{31}, E_{32}, E_{33}]^T \tag{8}$$

Stacking the constraints of five-point correspondences gives the linear equation (6) and by solving the system the parameters of E can be computed. Equations (6), (4), and (5) are extended to 10 cubic constraints, and then to a ten-degree polynomial. As a result, a maximum of 10 essential matrix solutions was obtained for any five-point set. The solution yielding the highest number of inliers was selected as a set representative. This five-point algorithm is applied in conjunction with preemptive RANSAC. A number of five-point sets are randomly taken from the total set of features. The five-point algorithm is applied to taken subsets and generate a number of hypotheses. The hypothesis with the best preemptive scoring which has the largest set of inliers is chosen as the final solution. This five-point algorithm may not always converge to a global minimum but can offer superior performance in the rotation because of some reasons:

- Essential matrix is estimated from a closed-form tenth degree polynomial.
- Five-point is a minimal set for essential estimation so the affection of outlier is small.
- Monocular method is not affected by imperfect calibration between left and right image of stereo camera.

Therefore, only left or right camera is used for rotation estimation.

3.2 Translation Estimation

The relative orientation between the previous and current frames was obtained by the algorithm described above. Here, we propose a joint forward-backward translation estimation. Firstly, an initial translation was estimated by one point RANSAC. Secondly, it is further improved by joint forward-backward non-iterative translation estimation with the rotation estimated previously.

Consider the projection equation from a 3D point feature from current frame to previous frame.

$$\begin{pmatrix} u_p \\ v_p \\ 1 \end{pmatrix} = \begin{pmatrix} f & 0 & u_c \\ 0 & f & v_c \\ 0 & 0 & 1 \end{pmatrix} \left[\begin{pmatrix} R_{3\times3} & t_{3\times1} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} - \begin{pmatrix} s \\ 0 \\ 0 \end{pmatrix} \right] \tag{9}$$

with:

- Homogenous image coordiate $(u_p, v_p, 1)^T$ in left or right frame of preivious frame.
- Focal length f.
- Rotation R and translation t from current frame to previous frame.
- 3D point (x, y, z) in current left frame.
- Value s is equal to 0 for left frame or baseline for right frame.

This projection equation is re-written detail as follows

$$\begin{cases} u_p &= f\dfrac{(x_{Rot} + t_x + s)}{z_{Rot} + t_z} + u_c \\ v_p &= f\dfrac{(y_{Rot} + t_y)}{z_{Rot} + t_z} + u_c \end{cases} \tag{10}$$

where

$$\begin{pmatrix} x_{Rot} \\ y_{Rot} \\ z_{Rot} \end{pmatrix} = R_{3\times3} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{11}$$

$$\begin{pmatrix} -1 & 0 & \dfrac{u - u_c}{f} \\ 0 & -1 & \dfrac{v - v_c}{f} \end{pmatrix} \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} = \begin{pmatrix} x_{Rot} + s - z_{Rot}\dfrac{(u - u_c)}{f} \\ y_{Rot} - z_{Rot}\dfrac{(v - v_c)}{f} \end{pmatrix} \tag{12}$$

A 3D feature of current frame is projected to both left and right of previous frame with $s = 0$ and $s = baseline$, respectively. So from Eq. (12), we can form a linear system of 4 equations of 3 unknown variables t_x, t_y, t_z as follows

$$A \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} = B \tag{13}$$

They are known to be solution of Pseudo Inverse method

$$\begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} = (A^T A)^{-1} A^T B \tag{14}$$

In idea case without feature noise, translation is successfully obtained by Eq. (14) using only one feature correspondence. However, in the real situation, feature noise is unavoidable, using one feature does not guarantee the success of estimation. To obtain higher accuracy of translation estimation, we wrap this algorithm into the RANSAC scheme, 100 samples of closest 3D features are used to estimate candidate translations. The best one producing the largest of number inliers is considered as the final solution. These inliers are further used for the refinement step. Different from conventional methods that minimize the re-projection error iteratively, our proposed refinement quickly estimates absolute translation by solving a linear system. In particular, all n inliers are plug into Eq. (13) to create

$$\begin{bmatrix} A_1 \\ A_2 \\ . \\ . \\ . \\ A_n \end{bmatrix} \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} = \begin{bmatrix} B_1 \\ B_2 \\ . \\ . \\ . \\ B_n \end{bmatrix} \tag{15}$$

Similar to above, Pseudo Inverse method is re-used to refine the initial estimation.

The above paragraph describes for backward estimation. To improve translation accuracy, both forward t_f and backward t_b translations are estimated by using same Eq. (15). The final solution is obtained by

$$t_{final} = 0.5(t_b - R^{-1} t_f) \tag{16}$$

where R is rotation estimated from previous section using the essential matrix.

4 Experimental Results

The proposed approach is evaluated on the KITTI dataset in comparison to its performance against the traditional VO pipeline, VISO2, proposed in [10]. The KITTI dataset consists of different traffic scenarios that are widely used

Table 1. Performance evaluation on KITTI Dataset

Sec Num	VISO2			Backward			Join For-Backward		
	t_e (%)	r_e (deg/ 100 m)	t_{abs} (m)	t_e (%)	r_e (deg/ 100 m)	t_{abs} (m)	t_e (%)	r_e (deg/ 100 m)	t_{abs} (m)
1	2.46	1.18	86.0	1.28	0.41	25.5	1.22	0.46	18.7
2	4.41	1.01	188.3	4.40	0.56	121.1	3.30	0.38	85.7
3	2.19	0.81	140.7	1.19	0.36	59.0	1.11	0.36	20.9
4	2.54	1.20	32.6	2.57	0.32	14.9	2.43	0.36	13.1
5	1.02	0.87	4.2	2.45	0.32	10.2	2.29	0.33	9.0
6	2.07	1.12	46.5	1.42	0.40	18.9	1.41	0.40	15.4
7	1.31	0.92	8.9	2.31	0.42	17.8	1.98	0.33	9.4
8	2.30	1.77	21.2	1.76	1.00	14.8	1.66	0.99	13.1
9	2.74	1.33	35.1	1.68	0.41	16.9	1.51	0.41	13.9
10	2.76	1.15	79.3	1.80	0.29	17.8	2.04	0.34	15.2
11	1.63	1.12	25.8	1.23	0.53	18.8	1.44	0.65	20.6
Avg	2.43	1.11	-	1.60	0.41	-	1.49	0.42	-

for evaluating autonomous driving algorithms. The dataset also accommodates challenging aspects such as different lighting, shadow conditions, and dynamic moving objects. In order to evaluate the performance of the VO approaches, RMSEs of measuring rotation/translation errors are computed from all possible sub-sequences of lengths (100, 200...800 m) as described in [14]. There is an evalution tool on their web page.

The detail error of 11 sections of dataset are shown in Table 1. For each approach, the Table displays the average rotation error r_e in degree/100 m, average translation in percentage (%) t_e and absolute error t_a in (m) of final frame compared to the ground-truth. The results of VISO2 library is named VISO2 and our proposed VO are named 'Backward' as well as 'Joint For-Backward', respectively, for two cases. This table indicates that the proposed approach achieves lower error for both translation and rotation, in general. Specifically, the rotation error of our approach using essential matrix is 0.46 deg/100 m while that of VISO2 is 1.1 deg/100 m; our translation error is 1.6% while that of VISO2 is 2.4%. That indicates around 30% translation error enhancement. The error of translation is further reduced to 1.49% by joint forward and backward translation estimation.That mean, the joint forward-backward estimation improve the translation accuracy 7%.

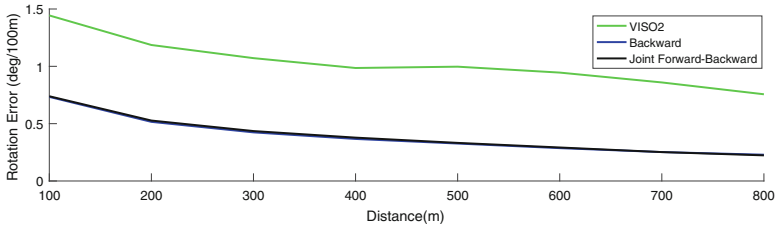

Fig. 2. Average rotation error along travel distance (Color figure online)

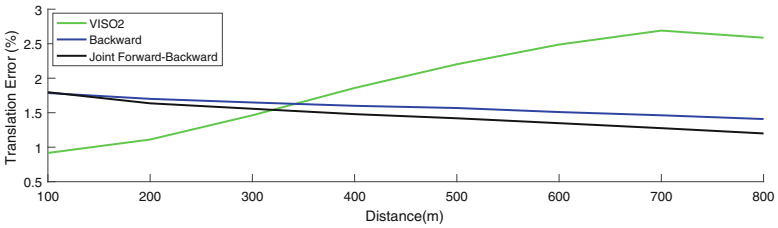

Fig. 3. Average translation error along travel distance

We also measure the transformation error along with the distance of travel from 100, 200,..., 800 m. The change of translation and rotation errors are shown in Fig. 2 and Fig. 3, respectively. For all travel distances, both rotation and translation errors of proposed approaches are smaller than those of VISO2. Specifically, Our translation error of backward estimation gradually reduces from 1.8% at 100 m to 1.5% at 800 m while the translation error of VISO2 increases monotonically from 1.0% at 100 m to 2.5% at 800 m. The change of joint forward-backward translation estimation in black is a little bit lower than that of backward estimation in blue. Consider the change of rotation error along the path length, our rotation errors are similar because we only focus on translation optimization. They gradually reduce from around 0.78 °/100 m at 100 m to around 0.23 °/100 m at 800 m. Similarly, The error of VISO2 reduces from 1.4 °/m at 100 m and 0.8 °/100 m at 800 m. However, at every travel distance 100, 200,..., 800 m. The error of the proposed approaches slower than that of VISO2.

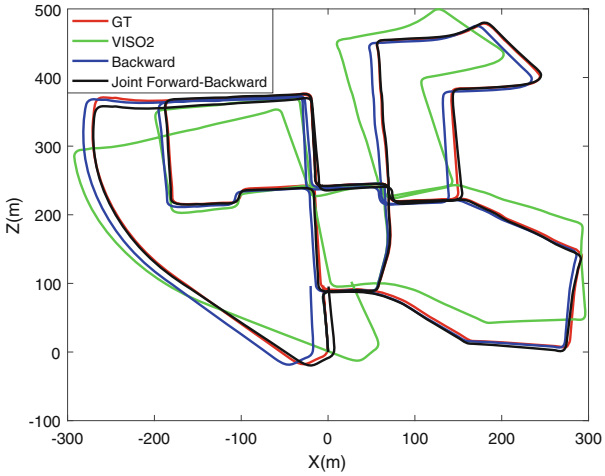

Fig. 4. Trajectory of Sect. 1 for three approaches compare to the ground-truth.

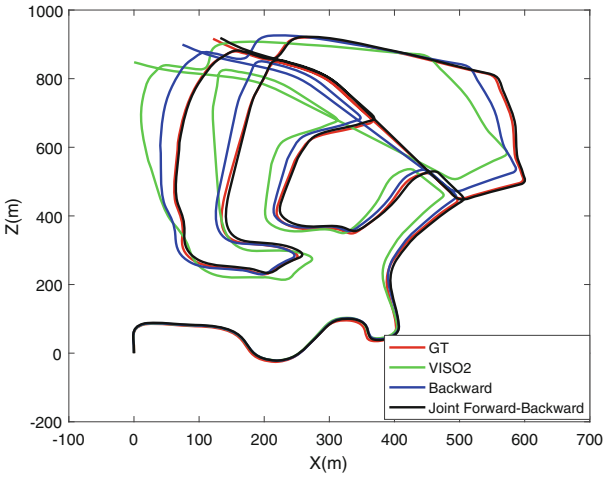

Fig. 5. Trajectory of Sect. 3 for three approaches compare to the ground-truth.

The accuracy improvement of our method compared to conventional approach VISO2 is confirmed by visualizing several camera trajectories in Sect. 1 and Sect. 3 in Fig. 4 and Fig. 5, respectively. It is clear that camera tracks of our approaches closer to the ground-truth than those of VISO2.

This evaluation has been done on an Intel Core i5-2400S CPU running at 2.5 Hz. The average single thread run-time per image for joint forward-backward translation estimation is 80 ms in total with 70 ms for rotation estimation and 10 ms for both forward and backward translation estimation. That means only forward or backward translation estimation spends on 5 ms.

5 Conclusion and Furture Work

An adaptive stereo visual odometry based on the essential matrix is presented by introducing the non-iterative translation estimation. The joint forward and backward translation estimation is proved to enhance performance. Compared to conventional methods that used PnP such as VISO2 library, the proposed method relies on the essential matrix estimation and the direct translation estimation by solving a linear system. The effectiveness of the proposed approach is verified by evaluating the errors in terms of translation and rotation on the KITTI dataset. The experimental result indicates that our approach achieves 1.6% and 1.49% with-out and with joint forward-backward translation estimation, respectively. In the future, we plan to widen the scope of applications utilizing the non-iterative translation estimation and refinement rotation estimation.

References

1. Nistér, D., Oleg ,N., James, B.: Visual odometry. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, vol. 1, pp. I-I. IEEE (2004)
2. Poddar, S., Kottath, R., Karar, V.: Motion estimation made easy: evolution and trends in visual odometry. In: Hassaballah, M., Hosny, K. (eds.) Recent Advances in Computer Vision. Studies in Computational Intelligence, vol. 804, pp. 305–331. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-03000-1_13
3. Scaramuzza, D., Friedrich, F.: Visual odometry [tutorial]. IEEE Robot. Autom. Mag. **18**(4), 80–92 (2011)
4. Friedrich, F., Scaramuzza, D.: Visual odometry: Part II: matching, robustness, optimization, and applications. IEEE Robot. Autom. Mag. **19**(2), 78–90 (2012)
5. Nistér, D.: An efficient solution to the five-point relative pose problem. IEEE Trans. Pattern Anal. Mach. Intell. **26**(6), 756–770 (2004)
6. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares fitting of two 3D point sets. IEEE Trans. Pattern Anal. Mach. Intell. **5**, 698–700 (1987)
7. Fischler, Martin A., Bolles, Robert C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
8. Longuet-Higgins, H.C.: A computer algorithm for reconstructing a scene from two projections. In: Fischler, M.A., Firschein, O. (eds.) Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, pp. 61–62. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1987)
9. Chli, M., Davison, A.J.: Active matching. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) European Conference on Computer Vision – ECCV 2008. Lecture Notes in Computer Science, vol. 5302, pp. 72–85. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_7
10. Kitt, B., Geiger, A., Lategahn, H.: Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In: 2010 IEEE on Intelligent Vehicles Symposium (IV), pp. 486–492. IEEE (2010)
11. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: dense 3D reconstruction in real-time. In: 2011 IEEE Intelligent Vehicles Symposium (IV). IEEE (2011)
12. Cvišić, I., Petrović, I.: Stereo odometry based on careful feature selection and tracking. In: 2015 European Conference on Mobile Robots (ECMR), pp. 1–6. IEEE (2015)
13. Nguyen, H.H., Lee, S.: Orthogonality index based optimal feature selection for visual Odometry. IEEE Access, p. 7 (2019)
14. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2012)