A novel translation estimation for essential matrix based stereo visual odometry

1st Huu-Hung Nguyen
Institute of System Integration
Le Quy Don Technical University
Hanoi, Vietnam
hungnh.isi@lqdtu.edu.vn(*)

2nd The-Tien Nguyen
Institute of System Integration
Le Quy Don Technical University
Hanoi, Vietnam
tiennt.isi@lqdtu.edu.vn

3th Cong-Manh Tran
Institute of System Integration
Le Quy Don Technical University
Hanoi, Vietnam
manh.isi@lqdtu.edu.vn

4rd Kim-Phuong Phung
Institute of System Integration
Le Quy Don Technical University
Hanoi, Vietnam
phuongpk.isi@lqdtu.edu.vn

5rd Quang-Thi Nguyen
Institute of System Integration
Le Quy Don Technical University
Hanoi, Vietnam
thinq.isi@lqdtu.edu.vn

***Abstract*—Visual Odometry (VO) plays an important role in autonomous navigation systems for vehicle localization. For traditional stereo visual odometry (SVO), we can estimate the rotation and translation of camera motion either simultaneously or separately where 3D information reconstructed from the stereo image is used as the input of the translation estimation. The accuracy of pose estimation is dependent on the uncertainty of 3D features as well as their portion used. This paper presents a novel translation estimation for essential matrix-based SVO to avoid the effectiveness of 3D feature uncertainty from stereo disparity. The rotation is extracted accurately from essential matrix of each pair of consecutive image frames on the left side; with a pre-estimated rotation matrix, the translation is rapidly and accurately estimated by solving a proposed linear closed-form only using 2D features as input with one-point RANSAC. The experimental results on the autonomous driving testing dataset (KITTI) indicate that the proposed approach enhances 20 % accuracy compared to traditional approaches in the same experimental scenario.**

***Index Terms*—Stereo Visual Odometry, Essential Matrix Estimation, Novel Translation Estimation**

I. Introduction

Localization and navigation play important role in an autonomous system. With rapidly advanced techniques in the field of mobile robotics, the requirement for accurate and efficient navigation and localization for an intelligent system has arisen. Camera-based localization is one of the most popular techniques due to its price and simplicity as well as resource limitation in generating motion-path. In general, this method determines the position and orientation of a robot by analyzing the associated camera images. It is so-called visual odometry (VO) that first was introduced by Moravec [1] and named by Nister in [2]. Recently, VO is classified into different approaches such as monocular/stereo camera-based, geometric/learning-based, and feature/appearance-based in the survey [3]. The feature-based VO pipeline has a long history and has been detailed in Nister's [2] work. Scaramuzza and Fraundorfer conducted a comprehensive review of feature-based VO [4], [5]. The feature-based VO classified into three approaches based on input data: 1) 2D-to-2D approach estimates the camera motion from 2D features only; 2) 3D-to-3D approach estimates camera motion from 3D features only 3) 3D-to-2D approach estimates motion from the 3D feature in one frame and corresponding 2D feature in other. The surveys in [2], [3] conclude that the 2D-to-2D and 3D-to-2D methods provide higher accuracy in pose estimation than 3D-to-3D one due to the uncertainty of the 3D feature. We can say that the more portion of 3D features using in the VO pipeline, the higher error in pose estimation compared to the ground-truth. The 3D-to-2D method is well-known as a perspective from n points (PnP). The pose is optimized via iteratively minimizing the summation of projection error between the 2D observations and projected points of corresponding 3D features.

The drift in trajectory over image frames is compensated using different strategies such as SLAM and Bundle Adjustment (BA) that aims at obtaining accurate motion vector given all the past feature positions and their tracking information [6]. Recently, several VO approaches reach the accuracy requirement without loop closure or bundle adjustment. VISO2 [7], for instance, is one of the most popular 3D-to-2D VO methods due to its efficiency and accuracy. This 3D-to-2D method is combined with the accuracy selection of keyframe and features to enhance the performance proposed in [9] (SSLAM). Essential matrix-based VO receives attention from many researchers

978-0-7381-0508-6/21/$31.00 ©2021 IEEE

due to the high accuracy in rotation estimation such as [13] and [12]. The Rotation matrix is initially extracted from the essential matrix and is finally refined by a loop constraint of 3 consecutive frames. Translation is initially estimated by different ways however it is finally refined by minimizing re-projection error with the pre-estimated rotation and using 3D features in one frame and corresponding 2D features in another frame as input. Their contributions lie in the feature selection technique in which 2D features are carefully selected for the pose estimation step.

We recognize that the usage of 3D features as an input of the translation estimation step may lead to highly erroneous estimation due to 3D uncertainties. This paper presents a novel translation estimation for essential matrix based stereo visual odometry. Different from the state-of-the-art methods using the more or less 3D information pre-calculated from disparity value for estimating translation, we investigate a linear closed-form expression in which translation and 3D information are calculated simultaneously. We validate our proposed idea on a publicly autonomous driving KITTI odometry dataset by comparing it to others. Our method achieves the average translation error around 1.17 %/m and rotation error 0.004 deg/m enhances 20% compared to the traditional method, re-projection minimization. Our algorithm is depicted in Fig 1

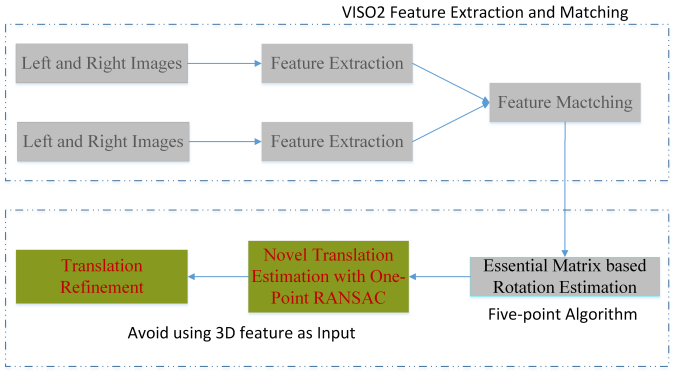

Fig. 1: The proposed VO pipeline with novel translation estimation

including two main phases as traditional approaches:feature extraction/matching and pose estimation. Our main contribution lies in highlighted as green block where translation estimation avoids using 3D information calculated from stereo disparity is proposed by estimating them simultaneously from 2D features.

The rest of this paper is organized as follows. Section II summarizes related works for essential matrix-based visual odometry. Section III deploys the novel translation estimation from the reprojection equation. The experimental results evaluating on the KITTI dataset are given in Section IV. Finally, Section V concludes this paper.

II. Related work

In this section, we summarize the essential matrix-based visual odometry. Usually, there are two main phases for pose estimation: rotation extraction from essential matrix estimation and translation estimation.

A. Rotation Estimation

The essential matrix $E_{3\times3}$ represents the geometric relation of a pair of consecutive image frames known as the camera motion parameters and is described in matrix form by equation (1)

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

where R is 3×3 rotation matrix and $T^{\times}$ is a 3×3 skew matrix generated from translation vector as follow

$$T^{\times} = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix} \tag{2}$$

Additionally, essential matrix E should satisfy two more internal geometric constraints as follows

$$det(E) = 0, \tag{3}$$

and

$$2EE^TE - tr(EE^T)E = 0 \tag{4}$$

Each pair of 2D correspondences between two image frames satisfies the epipolar constraint

$$p^TEq = 0 \tag{5}$$

Where a corresponding pair p, q are 2D features in the previous frame and current frame, respectively. Note that, essential matrix E is a 3×3 matrix including 3 unknown rotation parameters and 2 unknown translation parameters with an unobservable scale. Nister [10] proved that the essential matrix is possible to be solved by searching roots of tenth-degree polynomial expanding from the epipolar constraints of five correspondences and two internal constraints. This method is called five-point algorithm that is applied in conjunction with preemptive RANSAC [11] to choose the best solution with the minimum preemptive scoring and the largest number of inliers. This five-point algorithm may not always converge to the global minimum but can offer superior performance in the rotation because of some reasons:

- Solved by a closed-form tenth degree polynomial.
- Minimal noise affection due to using five-point as a minimal set for essential estimation.
- Avoid the imperfect stereo camera calibration between left and right image frames due to use the monocular method.

B. Translation Estimation

For essential matrix-based VO, essential matrix estimation is done by an efficient five-point algorithm proposed by Nister [10] in which a RANSAC scheme is used to choose the smallest preemptive score from N sets of five-point samples. Since the rotation is extracted directly from the estimated essential matrix, the missing information is three unknown parameters of the translation. The simplest solution is that use

a pair of 3D feature correspondence (P, Q) with the RANSAC scheme.

$$P = RQ + t, \tag{6}$$

where Q, P are 3D corresponding features of current and previous frames, respectively.

Similar to the 3D-to-3D method, this approach provides a translation solution with a high error due to the high uncertainty of 3D points. Recently, to avoid this uncertainty, the method in [12] estimated the translation by calculating the translation scale extracted from the essential matrix. However, conventional approaches use several techniques to initialize the parameters of translation, the final solution always is refined by minimizing the reprojection error (MRPE). In which the 3D-to-2D approach is used. The 3D feature in the current frame is projected to 2D image planes, the reprojection error defined by the difference of projected point and corresponding 2D observation.

$$RPE = \sum_{i=1}^{n} \left(w_1 \left(p_L - \mathbb{C}_{\mathbb{L}} Q \right)^2 + w_2 \left(p_R - \mathbb{C}_{\mathbb{R}} Q \right)^2 \right) \tag{7}$$

where Q is a 3D feature in the current frame that corresponds to two 2D features p_L and p_R in the previous frame. This approach has been applied successfully in [13] for the final optimization step. However, they control the weight $w1$, $w2$ by feature characteristics such as their age, strength...etc. To avoid the iterative process, the approach in [14] proposed an adaptive essential matrix-based stereo visual odometry with linear closed-form solution for both initial and final steps (AESVO). The target of this approach is also to minimize the reprojection error. Thank to the known rotation matrix $R_{3\times3}$, they transform the projection equation to the linear equation of translation parameters. They assume two frames involved previous and current frames.

- Previous frame have two left and right images L_1, R_1, respectively.
- Current frame have two left and right images L_2, R_2, respectively.

The transformation from left current to the left previous frame is expressed as:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}^{1L} = \begin{pmatrix} R_{11} & R_{12} & R_{13} & t_x \\ R_{21} & R_{22} & R_{23} & t_y \\ R_{31} & R_{32} & R_{33} & t_z \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}^{2L} \tag{8}$$

where

- Rotation $R_{3\times3}$, translation $t_{3\times1}$ from the current frame to the previous frame.
- 3D points $(X, Y, Z)^{1L}$, $(X, Y, Z)^{2L}$ in the previous and current frame, respectively.

Affine transformation is expressed by equation (9):

$$\begin{pmatrix} u_L \\ v_L \\ 1 \end{pmatrix} = \gamma \begin{pmatrix} f & 0 & c_u \\ 0 & f & c_v \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}^{1L} \tag{9}$$

with:

- Homogenous image coordinate $(u_L, v_L, 1)^T$ in left frame of previous frame.
- Focal length *f*.
- c_u, c_v are image center or principle point.

With the known rotation from essential matrix extraction, they combine equations (8) and (9) to convert the general form of projection to a linear equation of the translation described as equation 10

$$\begin{pmatrix} -1 & 0 & \frac{u - u_c}{f} \\ 0 & -1 & \frac{v - v_c}{f} \end{pmatrix} \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} = \begin{pmatrix} X_{Rot} + B - Z_{Rot} \frac{(u - u_c)}{f} \\ Y_{Rot} - Z_{Rot} \frac{(v - v_c)}{f} \end{pmatrix} \tag{10}$$

where

$$\begin{pmatrix} X_{Rot} \\ Y_{Rot} \\ Z_{Rot} \end{pmatrix} = R_{3\times3} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = R_{3\times3} Q$$

, B is baseline if the 2D projected point on the right side and $B = 0$ if the 2D projected point on the left side. However, the input of translation estimation still includes 3D features of the current frame and 2D features of the previous frame. We recognize that the traditional methods always using 3D features for translation estimation so that the 3D uncertainty more or less effects on the estimation. In this paper, we try to avoid using 3D features for translation estimation.

III. Optimal translation estimation

The orientation and translation are two typical components of pose estimation to identify camera motion. In this paper, we propose a novel method to improve the accuracy of the translation. Firstly, the value of rotation was obtained by using the five-point algorithm [10] described above part. Secondly, the translation was computed via the proposed equations without 3D input since projecting a 3D point in the current left frame (world coordinate) to the pixel coordinate of two consecutive frames. The proposed equations are described in detail below.

Different from [14], we want to avoid using 3D features as the input of translation estimation. Specifically, we combine (8) and (9) to transform the general projection formula from the world coordinate to pixel coordinate of the left previous frame into the linear equation of translation of a 3D point.

$$\begin{pmatrix} R_{31}\alpha_1 - R_{11} & R_{31}\alpha_2 - R_{21} \\ R_{32}\alpha_1 - R_{12} & R_{32}\alpha_2 - R_{22} \\ R_{33}\alpha_1 - R_{13} & R_{33}\alpha_2 - R_{23} \\ -1 & 0 \\ 0 & -1 \\ \alpha_1 & \alpha_2 \end{pmatrix}^T \begin{pmatrix} X \\ Y \\ Z \\ t_x \\ t_y \\ t_z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{11}$$

where

$$\alpha_1 = \frac{u_{1L} - u_c}{f}$$

$$\alpha_2 = \frac{v_{1L} - v_c}{f}$$

Similarly, projecting a 3D in the left current frame to the pixel coordinate of right previous frame can be expressed in equation (12)

$$\begin{pmatrix} R_{31}\alpha_3 - R_{11} & R_{31}\alpha_4 - R_{21} \\ R_{32}\alpha_3 - R_{12} & R_{32}\alpha_4 - R_{22} \\ R_{33}\alpha_3 - R_{23} & R_{33}\alpha_4 - R_{23} \\ -1 & 0 \\ 0 & -1 \\ \alpha_3 & \alpha_4 \end{pmatrix}^T \begin{pmatrix} X \\ Y \\ Z \\ t_x \\ t_y \\ t_z \end{pmatrix} = \begin{pmatrix} -B \\ 0 \end{pmatrix} \quad (12)$$

with, B is *stereo baseline* and

$$\alpha_3 = \frac{u_{1R} - u_c}{f}$$

$$\alpha_4 = \frac{v_{1R} - v_c}{f}$$

Subsequently, we implement projecting a 3D point of left current frame camera from the world coordinate to current left and right of frame camera of pixel coordinate, we get two equations as equation (11) and (12). In this case, the matrix of rotation $R_{3\times3}$ is identity matrix, and matrix of translation has formed: $\begin{pmatrix} t_x & t_y & t_z \end{pmatrix}^T = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^T$. Finally, for each feature correspondence, we get a system of linear equations as follows:

$$A_{8\times6} \begin{pmatrix} X \\ Y \\ Z \\ t_x \\ t_y \\ t_z \end{pmatrix}_{6\times1} = B_{8\times1} \quad (13)$$

The equation (13) includes 8 linear equations with 6 unknown variables. It can be solved via the Pseudo Inverse method to get value M. Note that M is a matrix 6×1 is calculated via the following formula:

$$M = \begin{pmatrix} X & Y & Z & t_x & t_y & t_z \end{pmatrix}^T = (A^T A)^{-1} A^T B \quad (14)$$

In an ideal case, the translation completely achieves by solving an equation (14) using only one feature correspondence. However, in real situations, the existing noise of features comes from different source such as light condition, imperfect camera calibration...Therefore, it is so difficult to get a good estimation when using only one feature. To guarantee accuracy of translation estimation, this algorithm is wrapped into the RANSAC scheme, with 100 samples of closest 3D features are used to estimate candidate translations. Finally, maximum inliers of best translation solution are used for refinement. In this paper, we propose a refinement method based on solving a linear system. The equation (13) is written for only one feature and it can be rewritten as the following equation:

$$\begin{pmatrix} A^{1XYZ}_{8\times3} & A^{1T}_{8\times3} \end{pmatrix}^1_{8\times6} \begin{pmatrix} X_1 \\ Y_1 \\ Z_1 \\ t_x \\ t_y \\ t_z \end{pmatrix}_{6\times1} = B^1_{8\times1} \quad (15)$$

in which, the matrix A is splited into 2 sub-matrices

$$A^1_{8\times6} = \begin{pmatrix} A^{1XYZ}_{8\times3} & A^{1T}_{8\times3} \end{pmatrix}$$

Generalizing equation (15) for N features, we will get an equation following:

$$A_n M_n = B_n \quad (16)$$

where

$$A_n = \begin{pmatrix} A^{1XYZ}_{8\times3} & 0 & ... & ... & A^{1T}_{8\times3} \\ 0 & A^{2XYZ}_{8\times3} & ... & ... & A^{2T}_{8\times3} \\ ... & ... & ... & ... & \\ 0 & 0 & ... & A^{nXYZ}_{8\times3} & A^{nT}_{8\times3} \end{pmatrix}_{8n\times(3n+3)}$$

$$M_n = \begin{pmatrix} X_1 & Y_1 & Z_1 & ... & X_n & Y_n & Z_n & t_x & t_y & t_z \end{pmatrix}^T_{1\times(3n+3)}$$

$$B_n = \begin{pmatrix} B_1 & B_2 & ... & B_n \end{pmatrix}^T_{1\times8n}$$

Similarly, the equation (16) is solved by the Pseudo Inverse method to refine the initial estimation. However in this case the size of the matrix A_n and B_n as well as the unknown M_n monotonously increase. Using a larger number of features suffers from high computational time. By several experimental tests also consolidate this guess. Moreover far distance features with small disparity do not provide a good contribution to enhance translation accuracy. To deal with this problem, we only choose 10 inliers closest 3D features with top largest disparity for refinement.

IV. Experiment results

In this paper, we used the KITTI dataset to evaluate our proposed approach. It is a publicly popular dataset that is created for evaluating autonomous driving algorithms based on vision. The dataset consists of different traffic scenarios that accommodate challenging aspects such as lighting, shadow conditions, and dynamic moving objects. There are 22 stereo sequences in total, saved in lossless png format. In which, the dataset is divided into 2 sub-sets for different purposes. Training dataset including eleven sequences (00-10) with ground-truth trajectories for offline performance evaluation. Testing dataset including remaining 11 sequences (11-21) without the ground-truth for online evaluation. We evaluate our approach on both two types of sub-sets. The performance of the VO approaches is based on the RMSEs of measuring rotation/translation errors. These metrics are defined in [15] by computing the average errors from all possible sub-sequences of lengths (100, 200,..., 800 meters). Our method compared to other methods as VISO2 [8], AESVO proposed in [14], MRPE based on the training dataset. Additionally, we show a comparison on testing dataset with VISO2, SSLAM [9].

A. Evaluation on Training Dataset

Firstly, both rotation and translation errors of 11 sections of training KITTI dataset are shown in more detail in Table I. It visualizes the average rotation error r_e in degree/100m, average translation in percentage (%) t_e and absolute error t_a in (m) between the final frame of estimation and the ground-truth.

TABLE I: Performance evaluation on KITTI Dataset

Sec Num	VISO2 [8]			AESVO_Backward [14]			MRPE [12]			Ours		
	t_e (%)	r_e ($\frac{deg}{100m}$)	t_{abs} (m)	t_e (%)	r_e ($\frac{deg}{100m}$)	t_{abs} (m)	t_e (%)	r_e ($\frac{deg}{100m}$)	t_{abs} (m)	t_e (%)	r_e ($\frac{deg}{100m}$)	t_{abs} (m)
1	2.46	1.18	86.0	1.28	**0.41**	25.5	1.11	0.46	18.2	**1.08**	0.46	**11.1**
2	4.41	1.01	188.3	4.40	0.56	121.1	8.20	0.42	180.6	**2.97**	**0.37**	**48.5**
3	2.19	0.81	140.7	1.19	**0.36**	59.0	**0.95**	**0.36**	39.3	0.98	0.40	**21.1**
4	2.54	1.20	32.6	2.57	**0.32**	14.9	**0.93**	0.45	6.8	1.05	0.40	**3.5**
5	1.02	0.87	4.2	2.45	0.32	10.2	0.66	**0.26**	**2.6**	**0.56**	0.34	3.4
6	2.07	1.12	46.5	1.42	0.40	18.9	0.88	0.40	17.7	**0.83**	**0.39**	**17.1**
7	1.31	0.92	8.9	2.31	0.42	17.8	1.11	0.49	18.9	**0.85**	**0.40**	**8.1**
8	2.30	1.77	21.2	1.76	**1.00**	14.8	3.23	1.56	27.6	**1.44**	1.28	**13.3**
9	2.74	1.33	35.1	1.68	0.41	16.9	1.32	0.42	19.7	**1.21**	**0.39**	**9.5**
10	2.76	1.15	79.3	1.80	0.29	17.8	**0.91**	**0.28**	**13.8**	1.24	0.36	16.9
11	1.63	1.12	25.8	1.23	**0.53**	18.8	**1.06**	**0.53**	**8.7**	1.61	0.68	19.1
Avg	2.43	1.11	-	1.60	**0.41**	-	1.44	0.43	-	**1.16**	0.43	-

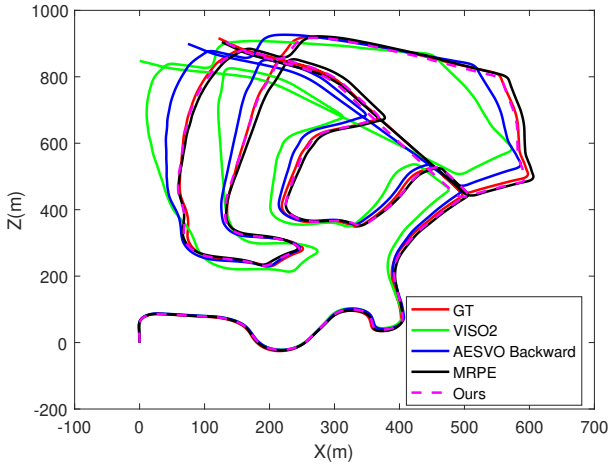

Fig. 2: Trajectory of sequence 2 for 3 approaches compare to the ground-truth

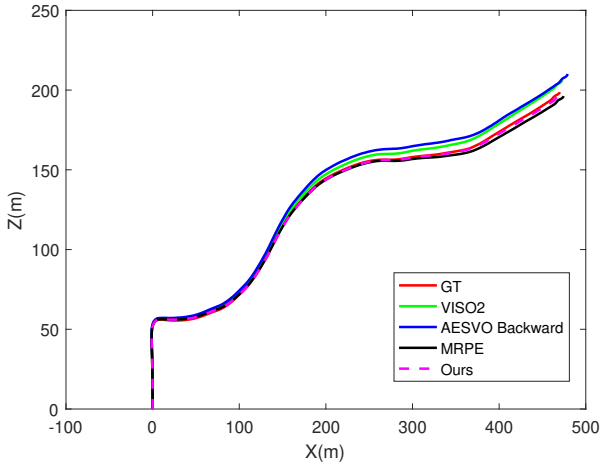

Fig. 3: Trajectory of sequence 3 for 3 approaches compare to the ground-truth

Table I shows the experimental results of 4 methods including the popular VISO2 [8], AESVO [14], MRPE [12], and the proposed method, respectively. To fairly compare approaches, the MRPE approach developed here without feature selection and rotation refinement by closed-loop of three consecutive frames.

Our proposed approach achieves lower errors for both translation and rotation in most of all sequences. Specifically, the average rotation error of VISO2, AESVO-Backward, and MRPE and ours is 1.11 deg/100m, 0.41 deg/100m, 0.43 deg/100m, and 0.43 deg/100m, respectively. Our rotation error is similar to that of AESVO and MRPE because of using the same essential matrix based approach. The accuracy of rotation estimation based essential matrix has been proved higher than that of 3D-to-2D method. The error of translation of VISO2, AESVO-Backward and MRPE are 2.43 %, 1.6%, and 1.44 %, respectively, while that of our proposed approach is 1.16 %.These results are understandable. The methods AESVO and MRPE still use 3D information as input so the high uncertainty of the 3D feature affects the accuracy of translation estimation. By avoiding using 3D information as input, our accuracy of translation is enhanced around 50%, 35%, and 20 % compared to VISO2, AESVO, and MRPE, respectively. These results validate that without using 3D as input can improve visual odometry accuracy.

Additionally, Fig.2 and Fig.3 illustrate the improvement of our proposed approach compared to the VISO2 approach by visualizing camera trajectories in sequence 2 and sequence 3. Look at these figures, we can realize that camera tracks of our proposed approach in pink closer to the ground-truth than others such as MRPE in black, AESVO in blue. These figures verify the accuracy of our method compared to others shown in Table I.

B. Evaluation on Testing Dataset

This sub-section shows the performance evaluation on the KITTI testing dataset that is fairly and publicly assessed on the web page. The result of our approach is compared to VISO2 and SSLAM [9] showed in Table II. The average of translation error of our approach is 1.42 (%) while the value of SSLAM

TABLE II: Performance evaluation on KITTI testing dataset

	Rotation Error (deg/100m)	Translation Error (%)
VISO2	0.0114	2.44
SSLAM	**0.0044**	1.57
Ours	0.048	**1.42**

[9] and VISO2 [8] are 1.57 (%) and 2.44 (%), respectively. It is clear that the translation error is improved than the VISO2 and SSLAM approaches.

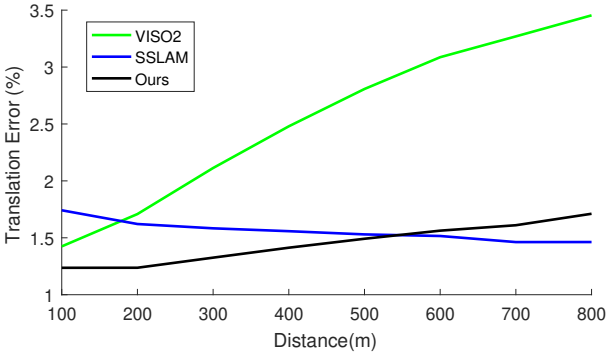

Fig. 4: Average translation error along travel distance

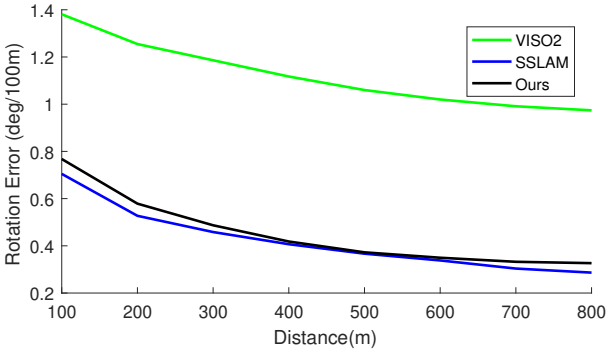

Fig. 5: Average rotation error along travel distance

We also measured the rotation and translation errors along with the distances. They are visualized in Fig.4 and Fig.5. Clearly, these errors of our approach are lower than that of VISO2. Compared to SSLAM, our rotation error is similar to the error of SSLAM. However, our translation error increase from 1.2 % at 100m to 1.7% at 800m while this error of SSLAM reduces from 1.8 % to 1.4 %. Finally, on average, our translation error is smaller than that of SSLAM and VISO2.

V. Conclusions and Future Work

A novel translation estimation without using 3D features as input for essential matrix-based visual odometry is presented. We investigate a simultaneous estimation of translation and 3D features with the known rotation extracted from the essential matrix. The accuracy of translation estimation is only dependent on the uncertainty of 2D features. The experimental results on the autonomous driving KITTI dataset prove that the proposed method enhances the accuracy of translation around 20% compared to traditional methods. In the future, we consider to further improve the performance by solving the existing issue: translation refinement using multiple features.

References

[1] H. P. Moravec, "The Stanford Cart and the CMU Rover," in Proceedings of the IEEE, vol. 71, no. 7, pp. 872-884, July 1983, doi: 10.1109/PROC.1983.12684.

[2] D. Nistér, N. Oleg and B. James, "Visual odometry." Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Vol. 1. Ieee, pp I-I, 2004.

[3] S. Poddar, R. Kottath, and V.Karar, "Motion Estimation Made Easy: Evolution and Trends in Visual Odometry". In Recent Advances in Computer Vision (pp. 305-331). Springer, Cham, 2019.

[4] D. Scaramuzza and F. Friedrich, "Visual odometry [tutorial]." IEEE robotics and automation magazine 18.4 (2011): 80-92.

[5] F. Friedrich and D. Scaramuzza. "Visual odometry: Part II: Matching, robustness, optimization, and applications." IEEE Robotics and Automation Magazine 19.2 (2012): 78-90.

[6] N. Fanani, A. Stürck, M. Ochs, H. Bradler, and R. Mester "Predictive monocular odometry (PMO): What is possible without RANSAC and multiframe bundle adjustment?". Image and Vision Computing, 68, 3-13, 2017.

[7] B. Kitt, A.Geiger, and H. Lategahn. "Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme." Intelligent Vehicles Symposium (IV), 2010 IEEE. IEEE, pp. 486-492, 2010.

[8] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time." Intelligent Vehicles Symposium (IV), 2011 IEEE. Ieee, 2011.

[9] F. Bellavia, M. Fanfani, F. Pazzaglia, and C. Colombo. Robust selective stereo SLAM without loop closure and bundle adjustment. In International Conference on Image Analysis and Processing (pp. 462-471). Springer, Berlin, Heidelberg, 2013.

[10] D. Nistér. "An efficient solution to the five-point relative pose problem." IEEE transactions on pattern analysis and machine intelligence 26.6 (2004): 756-770.

[11] Fischler, A. Martin , and C. Robert. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography." Communications of the ACM 24.6 (1981): 381-395.

[12] H. H. Nguyen and S. Lee, "Orthogonality Index Based Optimal Feature Selection for Visual Odometry," in IEEE Access, vol. 7, pp. 62284-62299, 2019, doi: 10.1109/ACCESS.2019.2916190

[13] I. Cvišić and I. Petrović, "Stereo odometry based on careful feature selection and tracking," 2015 European Conference on Mobile Robots (ECMR), Lincoln, 2015, pp. 1-6, doi: 10.1109/ECMR.2015.7324219.

[14] H. H. Nguyen, Q. T. Nguyen, C. M. Tran and D. S. Kim "Adaptive Essential Matrix based Stereo Visual Odometry with Joint Forward-Backward Translation Estimation" INISCOM2020, IEEE ,2020 [accepted]

[15] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[16] Our evaluation results on KITTI test dataset http://www.cvlibs.net/datasets/kitti/eval_odometry_detail.php?&result=aa1f1cf9c333f750ea8ed036a8cce56c38a04f15.