Fast Techniques for Monocular Visual Odometry

M. Hossein Mirabdollah(✉) and Bärbel Mertsching

GET Lab, University of Paderborn, Pohlwegstr 47-49, 33098 Paderborn, Germany
mirabdollah@get.upb.de

Abstract. In this paper, fast techniques are proposed to achieve real time and robust monocular visual odometry. We apply an iterative 5-point method to estimate instantaneous camera motion parameters in the context of a RANSAC algorithm to cope with outliers efficiently. In our method, landmarks are localized in space using a probabilistic triangulation method utilized to enhance the estimation of the last camera pose. The enhancement is performed by multiple observations of landmarks and minimization of a cost function consisting of epipolar geometry constraints for far landmarks and projective constraints for close landmarks. The performance of the proposed method is demonstrated through application to the challenging KITTI visual odometry dataset.

1 Introduction

Monocular visual odometry is known as a demanding problem in robotic and computer vision communities. The main challenge of a monocular odometry system is that feature depths are not measurable but rather they should be estimated. Unknown depths of features are mainly handled in literature based on two approaches. In the first approach, camera and feature positions are concurrently estimated in the context of extended Kalman filters. The methods belonging to this approach are mostly known as EKF-Monocular-SLAM methods (e.g. [3,9,16]). The main focus of this approach is how to parametrize large uncertainties of landmark positions in Gaussian forms in order to handle the problem in EKF filters. A good survey and comparison of these methods can be found in [15]. Among the EKF-based methods, the inverse depth parameterization (IDP) method [3] is known to be well established and has shown the best performance. However, it usually diverges if cameras move in depth. The reason is that this method localizes landmarks observed at low parallax angles very often behind cameras (negative depth problem). Additionally, complexity of the EKF based methods increases exponentially with respect to the number of landmarks, which makes them inappropriate for large scale robust visual odometry purposes. The second approach is based on bundle adjustment, in which a cost function between observed and predicted measurements (feature positions on the retina of a camera) at different camera poses is defined. Then the camera poses and feature positions are estimated by the minimization of the cost function. These methods require good initial guesses of camera poses. The initial guesses can be obtained from the epipolar geometry or based on the assumption

J. Gall et al. (Eds.): GCPR 2015, LNCS 9358, pp. 297–307, 2015.
DOI: 10.1007/978-3-319-24947-6_24

that the motion parameters of the camera do not change abruptly. Based on the epipolar geometry, a 3×3 matrix known as the essential matrix (for calibrated cameras) is estimated, which encodes camera motion. Essential matrices can be estimated using the 8-point [6], the 7-point [7] and the 5-point [13] methods.

In [19], the authors used bundle adjustment to minimize a cost function in which feature positions parametrized using IDP. This method is not real-time and may diverge if the camera moves in depth (due to the negative depth problem of the IDP). In [20], the 8-point method and a delayed parameterization technique known as the parallax angle parameterization are utilized to avoid the negative depth problem. This method essentially relies on the landmarks observed at high parallax angles. In [12], the authors used the perspective n point method (PnP) to estimate camera motion iteratively. The PnP method is mainly applicable if the positions of features in space are known (for instance from a stereo system). In case of a monocular system, it is assumed that the motion parameters do not change noticeably in consecutive frames; therefore, features can roughly be localized in space. Obviously, this method can only utilize features observed at high parallax angles and highly depends on the previous estimation of landmark positions. Hence, if the landmarks are not localized well in the previous steps, for instance due to measurement noise or small errors in the estimation of motion parameters, the method diverges. One common problem among the last three mentioned methods is that they cannot detect translation scale appropriately without using loop closure techniques. The reason is that visual features are hardly observed at high parallax angles in multiple frames. Consequently, the features cannot be used to detect scale drifts efficiently. Hence in the recent years, the scale detection problem has been approached in a different way. In case that a camera is installed on a wheeled vehicle and the height of the camera over the ground plane is known, absolute scale of camera motion can be determined. Geiger et al. in [5] used the 8-point method and the height of the camera over the ground plane to come up with the method known as libviso. Due to the usage of the 8-point method, libviso has a poor performance, especially in the estimation of rotation matrices. Additionally, in this method, they did not use any constraint to distinguish between the ground plane features from other features, resulting in large drifts in scale estimation. In [17,18], Song et al. developed multicore real time methods in which PnP is used to estimate motion parameters. Due to the usage of PnP, the methods produce large errors in case of bad localization of landmarks in previous steps. In another recent work proposed in [11], the 7-point method is modified to regularize roll and pitch angles of rotation matrices to enhance rotation estimations. This method is relatively time consuming and the rotation estimation is not as good as the PnP based methods.

In this paper, we propose a new visual odometry method which can handle far and close landmarks robustly. Our contribution to the monocular visual odometry is fourfold. First, using an iterative 5-point method to estimate initial guesses of motion parameters. Second, proposing a probabilistic triangulation method to obtain uncertainties of landmark positions. Third, robust tracking of low quality

features on ground planes to estimate scale of camera motion. Fourth, enhancing the last camera pose by minimization of a cost function containing epipolar and projective constraints to handle far and close landmarks intuitively. In our method, only camera poses are iteratively estimated and landmark positions are estimated based on the probabilistic triangulation method. This technique allows us to leverage hundreds of features in the optimization process in real time.

The paper is organized as follows: in Sect. 2, the iterative 5-point method is discussed. The probabilistic triangulation method is presented in Sect. 3. In Sect. 4, our method to detect scale of camera motion is proposed. Leverage of multiple observations of features is discussed in Sect. 5. The proposed algorithm is evaluated in Sect. 6. Section 7 concludes this paper.

2 Inter Frame Camera Motion Estimation

A typical approach to estimate camera motion parameters between two frames is using epipolar geometry. For a calibrated camera, given a set of matched points $\{(x, y), (x', y')\}$, the following equation holds:

$$[x'\, y'\, 1]E[x\, y\, 1]^T = 0 \tag{1}$$

where E is known as the essential matrix. Assuming that a coordinate frame is attached to each camera pose, each point in space in the first camera frame such as $\mathbf{p} = [p_x,\ p_y,\ p_z]^T$ will have the coordinate of $\mathbf{p}' = [p'_x,\ p'_y,\ p'_z]^T$ in the second frame obtained as follows:

$$\mathbf{p}' = R(\mathbf{p} - \mathbf{t}) \tag{2}$$

where R is a rotation matrix encoding the rotation from the second frame to the first frame and $\mathbf{t} = [t_x,\ t_y,\ t_z]^T$ is the translation of the second frame with respect to the first frame. It can be shown that the essential matrix is related to R and $\mathbf{t}$ as follows:

$$E = \begin{bmatrix} e_1 & e_2 & e_3 \\ e_4 & e_5 & e_6 \\ e_7 & e_8 & e_9 \end{bmatrix} = RT \tag{3}$$

where $T = [\mathbf{t}]_\times$ is an antisymmetric matrix.

As Nister discussed in [13], the 5-point method is the best algebraic method to estimate essential matrices. The good performance of the 5-point method stems from two facts: first, it deals with degenerate cases efficiently; second, it uses the minimal number of points to estimate essential matrices, which makes the 5-point method more robust against outliers in the context of a RANSAC algorithm [4]. Unfortunately, the 5-point method is complex and demanding to apply it for real time purposes. In [8], an iterative 5-point method is proposed, which runs in real time. Nevertheless, in this method, the possibility of more solutions is not considered and it delivers only one solution. Additionally, in this work, translation vectors are parametrized using two independent angles. This

parametrization produces more degree of nonlinearity and consequently more local minima in which the optimization process may get stuck. Here, we form a nonlinear equation system based on the Sampson distance [7] and two more constraints over the rotation and translation parameters. If we parametrize the rotation matrix with a quaternion $\mathbf{q} = [q_0, q_1, q_2, q_3]^T$, given five matched points such as $\{(x_i, y_i), (x'_i, y'_i)\}$, i = 1...5, the equation system will be:

$$\begin{aligned}
&\frac{\mathbf{e}^T \mathbf{f}_1}{\sqrt{a_1^2 + b_1^2 + a_1'^2 + b_1'^2}} = 0 \\
&\vdots \\
&\frac{\mathbf{e}^T \mathbf{f}_n}{\sqrt{a_5^2 + b_5^2 + a_5'^2 + b_5'^2}} = 0 \\
&q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1 \\
&t_x^2 + t_y^2 + t_z^2 = 1
\end{aligned} \tag{4}$$

where, $\mathbf{e} = [e_1, ..., e_9]^T$, $\mathbf{f}_i = [x'_i x_i, x'_i y_i, x'_i, y'_i x_i, y'_i y_i, y'_i, x_i, y_i, 1]^T$, $[a_i, b_i, c_i]^T = E[x_i, y_i, 1]^T$ and $[a'_i, b'_i, c'_i] = E^T[x'_i, y'_i, 1]^T$ (c and c' are not used in Eq. 4). The last two equations in Eq. 4 are due to the property of quaternions and the fact that the translation vector can only be estimated up to a scale factor. The above system of equations can be solved using the Gauss-Newton method. In iterative methods, initial guesses of parameters determine the converged solution. Thus, given five matched points, we obtain maximally up to 3 solutions based on the following initial guesses: $\mathbf{q} = [1, 0, 0, 0]^T$, $\mathbf{t} \in \{[1, 0, 0]^T, [0, 1, 0]^T, [0, 0, 1]^T\}$. Using the 5-point method in [13], we may obtain more solutions. However, in practical cases where rotations are not large, the other solutions are not either feasible or they are close to the solutions from the iterative method. Hence, the solutions are good enough to be used in the optimization process based on the multiple observations of landmarks.

3 Probabilistic Triangulation

We denote a camera pose at time t with respect to a global frame as $P_t = \{R_t, \mathbf{c}_t\}$, where R_t is a rotation matrix encoding the orientation of the camera and $\mathbf{c}_t$ shows the position of the camera in the global frame. If a landmark with the coordinate $\mathbf{p} = [p_x, p_y, p_z]^T$ is observed at two camera poses $P_k = \{R_k, \mathbf{c}_k\}$ and $P_t = \{R_t\ \mathbf{c}_t\}$ ($k < t$), at the points (x_k, y_k) and (x_t, y_t) on the retina of the camera, the landmark can be localized in space using triangulation. Our triangulation method is based on the fact that the point should lie on the lines drawn from the center of each camera pose in the directions of the observations. As a result, the following equations hold:

$$\mathbf{p} = \mathbf{c}_k + d_k \mathbf{v}_k \tag{5}$$

$$\mathbf{p} = \mathbf{c}_t + d_t \mathbf{v}_t \tag{6}$$

where $\mathbf{v}_k = R_k[x_k, y_k, 1]^T$ and $\mathbf{v}_t = R_t[x_t, y_t, 1]^T$. d_k and d_t are the depths of the landmark in the camera frames attached to each camera pose. Using the two equations, the following linear equation system is obtained:

$$[\mathbf{v}_k| - \mathbf{v}_t] \begin{bmatrix} d_k \\ d_t \end{bmatrix} = \mathbf{c}_t - \mathbf{c}_k = \mathbf{c}_{t,k} \tag{7}$$

By solving the above equation system, the depth of the landmark in the k^{th} camera frame will be:

$$d_k = \frac{\nu}{\rho} = \frac{\left(\mathbf{v}_t^T \mathbf{v}_t \mathbf{v}_k^T - \mathbf{v}_k^T \mathbf{v}_t \mathbf{v}_t^T\right) \mathbf{c}_{t,k}}{\mathbf{v}_t^T \mathbf{v}_t \mathbf{v}_k^T \mathbf{v}_k - (\mathbf{v}_k^T \mathbf{v}_t)^2} \tag{8}$$

It can be shown that if there are measurement noise or errors in the estimation of rotation and translation parameters, ρ and ν will be joint Gaussian random variables: $[\rho, \nu]^T \sim \mathcal{N}([\bar{\rho}, \bar{\nu}]^T, \Sigma)$. As a result, d_k has the distribution of the ratio of two dependent Gaussian random variables. It can be shown that:

$$p(d_k|\rho) = \frac{1}{\sqrt{2\pi}\sigma_\nu} \frac{|\rho|}{d_k^2} \exp\left(-\frac{(\rho - \bar{\nu} d_k)^2}{2\sigma_\nu^2 d_k^2}\right) \tag{9}$$

where σ_ν^2 is the variance of ν obtained from the marginalization of ρ from the joint distribution of ρ and ν. The goal of probabilistic triangulation is to find a confidence range for d_k such as $[d_k^{min}, d_k^{max}]$ at each new observation of the landmark. To this end, we use Eq. 9 for $\rho = \bar{\rho} - 2\sigma_\rho$ and $\rho = \bar{\rho} + 2\sigma_\rho$ and find two positive d_k at which the probability of $p(d_k|\rho)$ is equal to a small ratio of the maximum pick of the distribution. In Fig. 1, the two distributions for $\rho = 1$, $\sigma_\rho = 0.1$, $\nu = 0.1$ and $\sigma_\nu = 0.1$ are depicted.

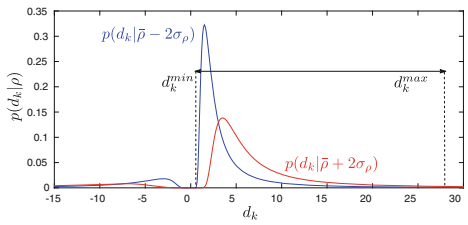

Fig. 1. Distribution of the depth parameter based on the probabilistic triangulation method.

It can be verified that the depth distribution tends to a Gaussian distribution in high parallax angles. In Eq. 8, the parallax angle is the angle between $\mathbf{v}_t$ and $\mathbf{v}_k$ ($\alpha = acos(\frac{\mathbf{v}_t^T \mathbf{v}_k}{|\mathbf{v}_t||||\mathbf{v}_k|||})$). We trim the range $[d_k^{min}, d_k^{max}]$ based on the new observations of the landmark such that $|d_k^{max} - d_k^{min}|$ reduces or stays the same. In another word, the uncertainty of a landmark position does not increase (in analogy to Bayesian filters) as the landmarks are assumed stationary.

4 Scale Detection

In case that a camera is installed on a wheeled vehicle parallelly to the ground plane, scale of translations can be obtained by using the height of the camera over the ground plane as a known parameter. Given R and $\mathbf{t}$ ($||\mathbf{t}|| = 1$) for two consecutive frames and the matched points $\{(x, y), (x', y')\}$, we use triangulation to localize the corresponding 3D point in the first camera frame as follows:

$$\mathbf{p} = d_1 \mathbf{v}_1 \tag{10}$$

where d_1 is the depth of the point in the first camera frame and $\mathbf{v}_1 = [x, y, 1]^T$. It can be shown that d_1 is linearly proportional to the scale factor: $d_1 = \eta s$. Thus, given the known height of the camera h, we have: $s = \frac{h}{y\eta}$.

To utilize the above mentioned method, it is required to track features on typically highly homogeneous ground planes. In this regard, we extract features at different resolutions from a rectangular region of interest in the half bottom of both images. Then for each feature in the first frame, we find two matches in the second frame based on the feature descriptor used in libviso [5]. An important criterion by which many of wrong matches can be filtered is the distances of the matched features to their corresponding epipolar lines. Using all of the matches, different scale factors are calculated and then by applying a median filter, the most probable scale factor is found. This method is fast and much more accurate than the the method used in libviso.

5 Multiple Observations of Landmarks

To deal with degenerate cases and also uncertainties of scale factors, multiple observations of landmarks should be leveraged. Hence, we optimize the current camera pose P_t based on the multiple observations of landmarks. To this end, we use two types of constraints: the epipolar constraint for landmarks observed at low parallax angles as their uncertainties are far from Gaussian distributions and the projective constraint for landmarks observed at high parallax angles. For a landmark observed for the first time at the camera pose $P_k = \{R_k, \mathbf{c}_k\}$ with the coordinate (x_k, y_k), the Sampson distance is defined as follows:

$$S_e = \frac{\mathbf{e}_{t,k}^T \mathbf{f}_{t,k}}{\sqrt{a^2 + b^2 + a'^2 + b'^2}} = 0 \tag{11}$$

where $\mathbf{e}_{t,k} = vect(R_{t,k}T_{t,k})$, $R_{t,k} = R_t^T R_k$, $T_{t,k} = [\mathbf{c}_t - \mathbf{c}_k]_\times$, $[a, b, c]^T = R_{t,k}T_{t,k}[x_k, y_k, 1]^T$, $[a', b', c']^T = T_{t,k}^T R_{t,k}^T [x_t, y_t, 1]^T$ and $\mathbf{f}_{t,k} = [x_t x_k, x_t y_k, x_t, y_t x_k, y_t y_k, y_t, x_k, y_k, 1]^T$.

In case of close landmarks, we can use the projective constraint:

$$(\mathbf{x}_t - \hat{\mathbf{x}}_t)^T M^{-1} (\mathbf{x}_t - \hat{\mathbf{x}}_t) = 0 \tag{12}$$

where $\mathbf{x}_t = [x_t, y_t]^T$ is the vector of the real measurement, $\hat{\mathbf{x}}_t = [\hat{x}_t, \hat{y}_t]^T$ is the vector of predicted measurement and M is a covariance matrix encoding the uncertainty of the measurement. $\hat{x}_t$ and $\hat{y}_t$ are calculated as follows:

$$\begin{aligned}\hat{x}_t &= \frac{\{R_t^T(\mathbf{c}_k + d_k R_k[x_k, y_k, 1]^T)\}_1}{\{R_t^T(\mathbf{c}_k + d_k R_k[x_k, y_k, 1]^T)\}_3} \\ \hat{y}_t &= \frac{\{R_t^T(\mathbf{c}_k + d_k R_k[x_k, y_k, 1]^T)\}_2}{\{R_t^T(\mathbf{c}_k + d_k R_k[x_k, y_k, 1]^T)\}_3}\end{aligned} \tag{13}$$

where $\{\}_i$ is the i^{th} element of a vector. In Eq. 12, M is calculated at each frame based on the uncertainty of the depth of the landmark. Hence to calculate M, we insert three samples: d_k, d_k^{min} and d_k^{max} in Eq. 13 and obtain three samples for the predicted measurement. Finally, based on the three samples, M is calculated. Now we can form a cost function which contains Sampson distances, projective constraints and a regularization constraint as follows:

$$C = \sum_{i=1}^{n_1} S_{e,i}^2 + \sum_{i=1}^{n_2} (\hat{\mathbf{x}}_{i,t} - \mathbf{x}_{i,t})^T M_i^{-1} (\hat{\mathbf{x}}_{i,t} - \mathbf{x}_{i,t}) + (\mathbf{y}_t - \hat{\mathbf{y}}_t)^T N^{-1} (\mathbf{y}_t - \hat{\mathbf{y}}_t) \tag{14}$$

where, n_1 and n_2 are the number of landmarks observed at low and high parallax angles respectively. $\mathbf{y}_t = [c_{x,t}, c_{y,t}, c_{z,t}, q_{0,t}, q_{1,t}, q_{2,t}, q_{3,t}]^T$ is a vector containing the parameters of the last camera pose. $\hat{\mathbf{y}}_t$ is the initial guess of the camera pose which is calculated based on the following motion model:

$$\begin{aligned}\mathbf{c}_t &= \mathbf{c}_{t-1} + R_{t-1}(\mathbf{q}_{t-1}) s\mathbf{t} \\ R_t(\mathbf{q}_t) &= R_{t-1}(\mathbf{q}_{t-1}) R^T\end{aligned} \tag{15}$$

where R and $\mathbf{t}$ are obtained from the inter frame camera motion estimation and s comes from the scale detection module. In Eq. 14, N is a covariance matrix obtained by the linearization of the motion model and error propagation through the linear model. In this regard, we consider some uncertainties for the instantaneous motion parameters. Experimentally, we found that the variance 0.0001 for the quaternion and translation elements works well. Additionally, the standard deviation of s is calculated dynamically based on the difference of two consecutive scale factors. The last term in the cost function is essential as the cost function could have several minima and the term regularizes the optimization process to converge to a state near to the initial guess (in the sense of Mahalanobis distances). The covariance matrix is also fed to the triangulation part, based on which the probabilistic triangulation is conducted. It should be mentioned that at each step the uncertainty of the previous camera pose is set to zero as we only use N as a regularization term in a smoothing scheme not a filtering scheme. In another word, we establish an intuitive relation between the unknown parameters and predefine the ranges of changes for each parameter in the optimization process. The overall method can be summarized as follows:

1. Given the last two images, calculate inter frame motion: R and $\mathbf{t}$.
2. Estimate scale of translation: s.
3. Predict the last camera pose and the covariance matrix N.
4. Minimize the cost function in Eq. 14. Use the Sampson distance for a landmark if $d_k^{max} - d_k^{min} > \Delta d_{threshold}$, otherwise use the projective constraint.
5. Run probabilistic triangulation.

6 Experimental Results

We implemented the proposed method in C++ and used the KITTI visual odometry dataset for the evaluation. Concerning feature tracking, Shi-Thomasi corner features [14] with the minimum quality of 0.01 were extracted and tracked using the Lucas-Kanade optical flow method (LK) [10]. Both of the algorithms are implemented in OpenCV [2]. The minimum distance between features was 30 pixels and the maximum number of features was 300. For the estimation of motion parameters between two frames, the iterative 5-point method discussed in Sect. 2 was used. The parameters were updated in fixed number of 5 iterations. The features were tracked maximally within 10 frames and $\Delta d_{threshold} = 15$. Based on multiple observations of features, the cost function in Eq. 14 was optimized with 5 iterations. With this setup, we achieved a real time performance (10 Hz) on a PC with an Intel Xeon E31270 @ 3.40GHz CPU without using any parallelism technique. For the evaluation, two measures are used: translation and rotation errors. Given the real position of a camera at time t as $\mathbf{c}_t$ and the estimated camera position as $\hat{\mathbf{c}}_t$, the average translation error is calculated as follows:

$$\epsilon_{\mathbf{c}} = \frac{1}{N_f} \sum_{t=0}^{N_f-1} \|\mathbf{c}_t - \hat{\mathbf{c}}_t\| \tag{16}$$

where N_f is the number of frames. The average rotation error is defined as:

$$\epsilon_R = \frac{180}{\pi N_f} \sum_{t=0}^{N_f-1} \left| acos\left(\frac{trace(R_t^e) - 1}{2}\right)\right| \tag{17}$$

where $R_t^e = R_t^T \hat{R}_t$. We compared our method based on the multiple observations (MO) and only two view optimization using our iterative 5-point method (TVO) with libviso (LV) [5], the iterative method in [8] (I5p) and a visual odometry method based on the normalized 8-point method [6] and the LK tracker (8pLK). In Table 1, the translation and rotation errors for some of the challenging training sequences of the KITTI dataset and also the average errors for all 11 sequences are presented. Interestingly, we see that only applying our iterative 5-point method (TVO) yields dramatically better estimations in comparison to the other two view based methods. In average, I5p has the poorest performance as it neglects possibility of multiple solutions and also gets stuck in local minima due to the way it parametrizes the essential matrix. Especially, it performs poorly for sequences where the car often drives through sharp bends

(due to the occurrence of degenerate cases). Interestingly, 8pLK performs better than libviso, which signifies superiority of LK tracker over the feature matching technique used in libviso as LK provides sub-pixel accuracies resulting in less measurement noise. As expected, the multiple view observation technique enhances the results from TVO, especially for the sequences where the ratio of outliers is high or the number of observed features at high parallax angles is low (for instance sequence 1). In Fig. 2, the estimated paths for the sequence 1 using MO, TVO and LV are visualized. In this sequence, the car drives in an autobahn and the number of landmarks observed at high parallax angles is low. As can be seen, TVO has a poor performance when estimating the elevation of the camera (originated from the error in the estimation of roll and pitch angles); whereas MO is able to estimate the path well.

Table 1. Average of translation and rotation errors using different methods for the training sequences of KITTI dataset.

Seq.	N_f	Method	MO	TVO	8pLK	LV	I5p	MO	TVO	8pLK	LV	I5p
		Length [m]	$\epsilon_{\mathbf{c}}$ [m]					ϵ_R [deg]				
0	4541	3723.6	**10.4**	29.6	65.5	283.2	129.2	**1.4**	2.1	32.4	43.2	37.7
1	1101	2453.1	**97.9**	171.7	495.7	867.0	312.7	**4.7**	7.1	49.1	50.15	13.3
2	4661	5067.0	**32.3**	39.9	63.9	229.5	491.9	**1.2**	1.5	5.8	17.6	39.1
7	1101	694.7	**25.7**	89.6	123.3	115.1	99.3	**2.6**	3.7	4.3	40.9	22.1
Avg.	2109.2	2016.1	**21.3**	38.6	83.2	224.0	233.1	**1.9**	2.8	8.7	22.9	31.8

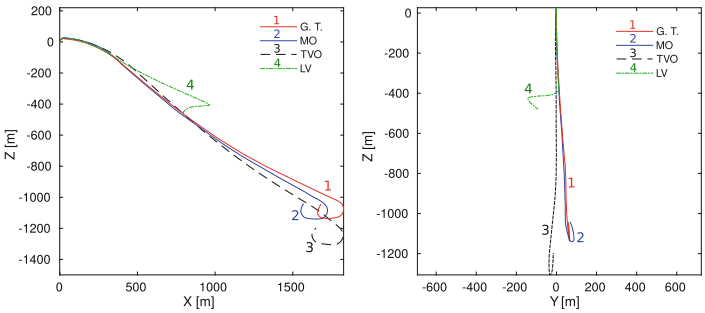

Fig. 2. Ground truth (G. T.) and estimated paths using different methods for the sequence 1 of the KITTI dataset.

We also submitted our results for the test sequences to the KITTI website under the name of FTMVO [1]. In the KITTI website, the methods are evaluated based on the percentage of errors until 800 meters with the step of 100 meters. In Table 2, the average of translation and rotation errors for our method and two recent methods of state-of-the-art are presented. As can be seen, our method outperforms the two methods MLM-SFM [17] and RCMPE+GP [11]. In [1], it can be seen that our method also outperforms many of the stereo vision based

Table 2. Average of translation and rotation errors for the test sequences of KITTI dataset: our method (FTMVO), MLM-SFM (M. 1) and RCMPE+GP (M. 2).

Method	FTMVO	M. 1	M. 2
Tr. error [%]	2.24	2.54	2.55

Method	FTMVO	M. 1	M. 2
Rot. error [deg/m]	0.049	0.057	0.087

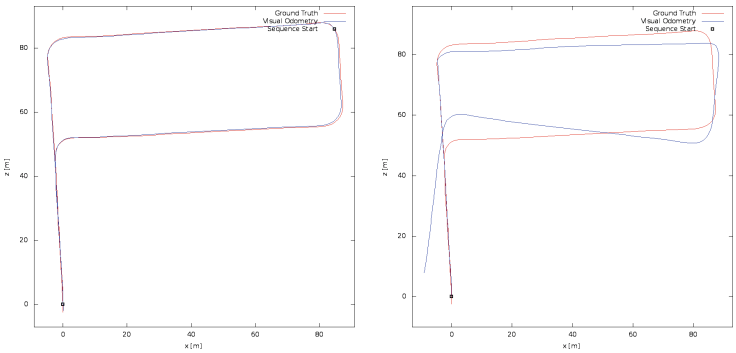

Fig. 3. Estimated (blue) and ground truth (red) paths for test sequence 14 based on our method (left) and MLM-SFM (right) (Colour figure online).

methods. From the test sequences, the $X - Z$ path of the first five sequences are visualized in the KITTI website. In Fig. 3, the estimated paths using our method and MLM-SFM for the sequence 14 are shown. The poor performance of MLM-SFM for this sequence lies in using the PnP method which degrades the estimations if the landmarks are badly localized in the previous frames. This situation occurs often if the camera experiences relatively large rotations and small translations.

7 Conclusion

An intuitive monocular visual odometry method is proposed, in which far and close landmarks are robustly handled. Through the proposed probabilistic triangulation technique, unlike the common SLAM or structure from motion methods, we can run the optimization process only on the last camera pose and exclude the localization of landmarks from the optimization process. Such an approach results in speeding up the algorithm to a great extent and also robustness of the algorithm against outliers. The performance of the method is demonstrated based on the large and demanding KITTI dataset for visual odometry.

References

1. Kitti visual odometry data set. http://www.cvlibs.net/datasets/kitti/eval_odometry.php
2. Bradski, G.: OpenCv library. Dr. Dobb's J. Softw. Tools **25**(11), 120–126 (2000)

3. Civera, J., Davison, A., Montiel, J.: Inverse depth parametrization for monocular SLAM. IEEE Trans. Robot. **24**(5), 932–945 (2008)
4. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
5. Geiger, A., Ziegler, J., Stiller, C.: StereoScan: Dense 3d reconstruction in real-time. In: Proceeding of Intelligent Vehicles Symposium (2011)
6. Hartley, R.: In defense of the eight-point algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **19**(6), 580–593 (1997)
7. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004). (ISBN: 0521540518)
8. Hedborg, B., Felsberg, M.: Fast iterative five point relative pose estimation. In: Proceeding of IEEE Workshop on Robot Vision, pp. 60–67 (2013)
9. Kwok, N.M., Dissanayake, G., Ha, Q.: Bearing-only slam using a SPRT based Gaussian sum filter. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 1109–1114 (2006)
10. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 674–679 (1981)
11. Mirabdollah, M.H., Mertsching, B.: On the second order statistics of essential matrix elements. In: Jiang, X., Hornegger, J., Koch, R. (eds.) GCPR 2014. LNCS, vol. 8753, pp. 547–557. Springer, Heidelberg (2014)
12. Mur-Artal, R., Tardos, J.: ORB-SLAM: tracking and mapping recognizable features. In: Proceeding of Robotics: Science and Systems (RSS) Workshop on Multi View Geometry in Robotics (2014)
13. Nistér, D.: An Efficient solution to the five-point relative pose problem. IEEE Trans. Pattern Anal. Mach. Intell. **26**(6), 756–777 (2004)
14. Shi, J., Tomasi, C.: Good features to track. Technical report (1993)
15. Solà, J., Vidal-Calleja, T., Civera, J., Montiel, L.M.: Impact of landmark parametrization on monocular EKF-SLAM with points and lines. Int. J. Comput. Vis. **97**(3), 339–368 (2012)
16. Sola, J., Monin, A., Devy, M., Lemaire, T.: Undelayed initialization in bearing-only SLAM. In: Proceedings IEEE International Conference on Intelligent Robots and Systems, pp. 2499–2504 (2005)
17. Song, S., Chandraker, M.: Robust scale estimation in real-time monocular SFM for autonomous driving. In: Proceeding of Computer Vision and Pattern Recognition (2014)
18. Song, S., Chandraker, M., Guest, C.: Parallel, real-time monocular visual odometry. In: Proceeding of IEEE International Conference on Robotics and Automation, pp. 4698–4705 (2013)
19. Strasdat, H., Montiel, J.M.M., Davison, A.: Scale drift-aware large scale monocular SLAM. In: Proceedings of Robotics: Science and Systems (2010)
20. Zhao, L., Huang, S., Yan, L., Dissanayake, G.: Parallax angle parametrization for monocular SLAM. In: Proceeding of IEEE International Conference on Robotics and Automation, pp. 3117–3124 (2011)