Absolute Scale Estimation Approach for Monocular Visual Odometry

Aldo André Díaz–Salazar
Federal University of Goias (UFG–Brazil)
Institute of Informatics (INF)
aldo.diaz@ufg.br

Paulo Roberto Gardel Kurka
University of Campinas (UNICAMP–Brazil)
School of Mechanical Engineering (FEM)
kurka@fem.unicamp.br

Abstract

Monocular visual odometry is an effective motion estimation technique that requires to solve for the challenging problem of absolute (metric) scale estimation. Current approaches use information such as the camera height or size of known objects to estimate the scene scale. In this paper, we propose a novel prediction-correction method to estimate the absolute scale of motion using camera height and flat ground assumption. Prediction is provided by a robust relative scale estimation strategy that exploits redundancy in depth information. Correction implements ground patch correlation using subpixel search refinement. The proposed method is tested using the public KITTI benchmark. As result, we derive analytical expressions to determine the absolute scale using a monocular camera. The empirical results shows the effectiveness of the proposed absolute scale estimation strategy in reducing the scale drift in monocular visual odometry.

1. Introduction

Visual odometry (VO) uses cameras as perception sensors for motion estimation which have important characteristics such as low cost hardware and rich source of information contained in camera images (color, semantic content and geometry). Recent applications have been enabled due to recent advances in this field, such as augmented reality, aerial navigation and autonomous vehicles. The automotive industry has particularly driven research and development on visual odometry methods due to is capability of delivering commercial products for autonomous driving applications. In this context, the methods made use of the particular characteristics of the environment for automotive vehicle navigation, such as landscape features and object detection (pedestrians, cars, trees), physical mounting constraints (fixed camera installation), lane lines and road features and signs. There is a number of publicly available datasets that provide benchmark data for the evaluation of visual odometry algorithms. Among all, the KITTI dataset is currently one of the most popular datasets due to the reference metrics they provide in open access to enable quantitative comparison of results. Regarding the camera configuration, the two main approaches use stereo camera or single monocular camera. Stereo configurations provide motion estimates with absolute scale as the transformation between cameras is known, but they require extra processing effort to compute the information of an additional camera. In contrast, odometry based on monocular camera, which has been argued to be a more challenging problem [7], has demonstrated similar efficiency to the stereo case [8]. Estimating the absolute scale of motion is a key challenge in moncular odometry methods.

Several strategies addressing absolute scale estimation in monocular odometry are reported in literature. We report a short summary of the methods with reported results on the KITTI dataset for the sake of comparison using a publicly available benchmark. The methods can be classified in three main categories.

Multi-sensor methods use heterogeneous sensors such as inertial, cameras or lidars. In [6], it was developed a general SLAM system called ORB-SLAM2 for monocular, stereo and RGB-D cameras including features such as map reuse, loop closing and relocalization. Motion estimation is provided by bundle adjustment using stereo observations to retrieve the absolute scale. The system achieved real-time operation in a variety of different environments (indoors, outdoors) with $1.15\,\%$ of MRPE error on KITTI dataset.

Camera mounting methods use physical parameters of the sensor system as sources of metric information. In [5], it were used monocular techniques such as the 5-point algorithm for estimating camera motion. The camera height is used to track low quality features on the ground plane with a robust approach. They achieved $2.24\,\%$ of relative translation error in the KITTI dataset.

Learning based methods use artificial intelligence pipelines. In [9], it was used deep learning to predict point depths of monocular vision. The method applies a supervised training stage that based on sparse depth reconstruction from stereo images using *direct sparse odome-*

try (DSO). In the training step, depth reconstruction applying DSO to stereo images is used to predict the monocular depths. It was obtained a comparable performance to stereo methods on KITTI dataset using a single camera with $0.90\,\%$ of translation error.

The monocular visual odometry system we propose uses camera height information and flat ground assumptions only. We provide a procedure that can be generalized to ground odometry scenarios whenever former two conditions are fulfilled. Compared to related methods based on camera height, we developed a prediction-correction mechanism to determine the absolute scale using a robust relative scale estimation step that exploits redundancy of keypoint depth information. We also derive an analytical expression to determine the absolute scale from reconstructed depths of 3-D points located on the ground.

2. Methodology

We proposed a two-step prediction-correction strategy based on the following assumptions:

1. **Relative scale estimation (prediction):** Unstructured outdoor environment. Motion estimation of a monocular camera is recovered by processing natural visual features present on the environment up to a relative scale and unknown absolute scale factor. Motion estimates are rotation $\mathcal{R}$ and translation $\boldsymbol{t}$.

2. **Absolute scale estimation (correction):** Camera is mounted rigidly at a known height; the ground floor is assumed flat in the local neighborhood closest to the camera field of view. We made use of the specific constraints of the system such as the parameters of camera mounting (height and orientation) and the flat ground assumption in order to determine the absolute scale of camera motion.

2.1. Scale detection mechanism

The calculation of the absolute scale of the translation uses camera height and ground plane information by assuming a flat world. Fig. 1 depicts a camera mounted at a height h and orientation $\mathcal{R}_m$ in forward looking position (pointing towards the direction of forward motion). The coordinate systems are denoted w for the world frame and c_k for the camera frame at instant k. The ground plane is represented by a point in the ground $\boldsymbol{P}_g^w$.

With the information of the camera height and orientation it is possible to recover the absolute scale of translation by relating a point in the ground in two subsequent images and solving for the 3-D structure using the planar geometry assumption. We first choose a patch on the ground plane by defining a *region of interest* (ROI), which is a small square image of the ground. *Backtracking* is then performed which consists of finding the ROI image correspondence in the image of a previous instant using the information of the relative scale. Backtracking provides a prediction of the possible ROI position in the image. The ROI position is later refined by searching for the best patch correlation within a local pixel neighborhood. The region of maximum correlation is used as the ROI correspondence. The 3-D depths of the two ROI correspondences is calculated from the homography relation defined using the road plane assumption. Finally, the absolute magnitude of the translation vector is recover by triangulation.

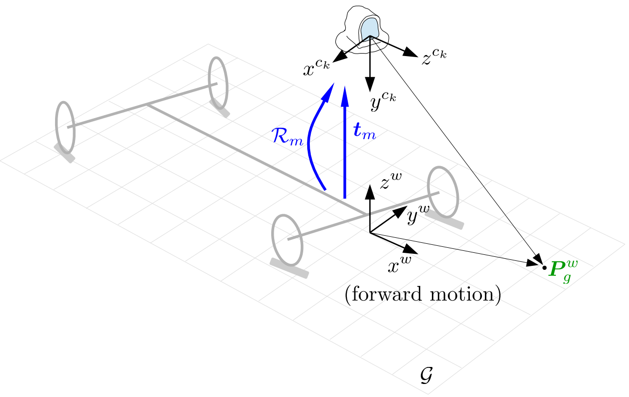

Figure 1: Forward looking camera mounted rigidly on the vehicle. The camera mounting parameters are the camera height $\boldsymbol{t}_m = [0, 0, h]^T$ and the camera orientation $\mathcal{R}_m$. Part of the ground plane lies within the field of view of the camera (note the point on the ground of coordinates $\boldsymbol{P}_g^w = [x^w, y^w, 0]^T$) which is valuable information to correct for the absolute scale of translational motion using the plane homography method.

2.2. Absolute scale estimation

The translation vector is expressed in terms of a scale parameter and a unit vector

$$\boldsymbol{t}_k = \lambda \hat{\boldsymbol{t}}_k\,, \tag{1a}$$

$$\hat{\boldsymbol{t}}_k = \frac{\boldsymbol{t}_k}{\|\boldsymbol{t}_k\|}\,. \tag{1b}$$

Now we can derived an expression for the absolute scale, denoted λ, by relating a reconstructed 3-D point in the ground from consecutive camera frames c_{k-1} and c_k, such as

$$\boldsymbol{P}_g^{c_{k-1}} = \mathcal{R}_k^T \boldsymbol{P}_g^{c_k} + \boldsymbol{t}_k\,, \tag{2a}$$

$$\boldsymbol{P}_g^{c_{k-1}} = \mathcal{R}_k^T \boldsymbol{P}_g^{c_k} + \lambda \hat{\boldsymbol{t}}_k\,, \tag{2b}$$

$$\lambda = \hat{\boldsymbol{t}}_k^T (\boldsymbol{P}_g^{c_{k-1}} - \mathcal{R}_k^T \boldsymbol{P}_g^{c_k})\,. \tag{2c}$$

Note that given motion parameters $\mathcal{R}_k$ and $\hat{\boldsymbol{t}}_k$ are estimated from visual odometry using the relative scale estimation al-

Algorithm 1 Absolute scale estimation using plane homography.

```
Given the position of the ROI center at current instant $k$.
1: Calculate the predicted position of ROI at instant $k-1$ (backtracking) using the information of the relative scale.
2: Correct the predicted ROI position by finding the patch of maximum correlation on a local patch neighborhood.
3: Recover the depths and 3-D coordinates of the point in the ground at the two instants $k-1$ and $k$.
4: Estimate the absolute scale of the translation from (2).
```

gorithm, as explained previously (see Fig. 1). The procedure is summarized in Algorithm 1.

3. Results

The results of Table 1 show the absolute scale estimation results using the KITTI evaluation benchmark. The environment included urban, highway, and countryside scenes of different lengths (**L**). The KITTI evaluation metrics measure the average translation ($\boldsymbol{t}$) and rotation ($\mathcal{R}$) errors by separate as a function of the trajectory length and velocity to provide means for error control over time [3], and are standard metrics for the evaluation of visual odomery algorithms [2]. The translational and rotational errors are calculated using the average of different trajectory lengths at ($100\,\mathrm{m}, 200\,\mathrm{m}, \dots, 800\,\mathrm{m}$), where errors are measured in percent for translation and in degrees per meter for rotation. The results show proper estimated trajectories with average translation error below $16\,\%$ for 6 out of 10 sequences (00, 02, 03, 04, 05, 06, 08). This sequences are characterized by an initial forward motion with sufficient parallax (large baseline) that enables an accurate initial absolute scale. The sequences 00, 02, 06 and 08 are characterized by more aggressive curves on rotations. The absolute scale corrects for motion estimates even in the presence of several rotations.

3.1. Computational complexity

Performance tests were calculated using sequence 00 as reference. The average computational time per iteration of our absolute scale estimation algorithm for a Python implementation run on a 2.3 GHz Intel Core i5-6200U processor was 2.3553 seconds. Note that our ROI correspondence mechanism has not been optimized for computational speed and current implementation takes an average of 1.9858 seconds per iteration which is about $84\,\%$ of the total computation time. The average performance of the relative scale estimation algorithm was 0.2453 seconds per iteration.

Table 1: Absolute scale estimation results on KITTI datasets. Errors are measured using trajectory segments at $100\,\mathrm{m}, 200\,\mathrm{m}, \dots, 800\,\mathrm{m}$, as an average of segment lengths $\%$.

ID	**L** (m)	**Environment**	$\boldsymbol{t}$ (%)	$\mathcal{R}$ (deg /m)
00	3714	Urban	11.72	0.0007
01	4268	Highway	79.00	0.0003
02	5075	Urban+Country	15.73	0.0006
03	563	Country	15.80	0.0003
04	397	Country	2.70	0.0001
05	2223	Urban	10.46	0.0004
06	1239	Urban	4.60	0.0002
07	695	Urban	37.24	0.0006
08	3225	Urban+Country	12.24	0.0006
09	1717	Urban+Country	43.93	0.0005
10	919	Urban+Country	61.86	0.0008

Optimization steps can be taken in order to speed up the current implementation. For instance, using a dedicated processor for computing ROI correspondence (*e.g.*, ASIC, FPGA, GPU). In this sense, our method is prone of such hardware optimization due to its modular design and, therefore, real-time implementations are feasible.

4. Analysis

Our visual odometry approach was to push the limits of using a single camera with a minimum set of assumptions and prerequisites (no learning, extra sensors or offline processing were introduced). Our algorithm achieved 20.08% on average translational error in the KITTI ranking against state-of-the art results of up to 21.47% [2]. Compared to related monocular approaches reported in [2], in [1] it was achieved 2.05% on average translational error, and [5] obtained 2.24%. The main difference is that those methods are based on sources of external information, such as experienced-based (machine learning), extra sensors (lidar, inercial, stereo camera), offline processing (pose-refinement, batch-processing), or any combination of them. This may explain the difference in accuracy.

5. Conclusion

We have developed a novel absolute scale estimation algorithm for monocular visual odometry using the camera height information and the flat ground assumption. An analytical expression for the absolute scale of translation is derived based on a fixed camera height. The proposed algorithm achieved state-of-the-art ranking in KITTI dataset with an average error of $20.08\,\%$ for translation and $0.0006\,\mathrm{deg}\,/\mathrm{m}$ for rotation. For future work, a velocity analysis mechanism can be used to overcome wrong scale estimates in small baseline motion.

References

[1] Nolang Fanani, Alina Sturck, Marc Barnada, and Rudolf Mester. Multimodal scale estimation for monocular visual odometry. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1714–1721. IEEE, jun 2017. 3

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Visual Odometry / SLAM Evaluation 2012, 2021. 3

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, jun 2012. 3

[4] Yi Ma, Stefano Soatto, Jana Košecká, and S. Shankar Sastry. *An Invitation to 3-D Vision*, volume 26 of *Interdisciplinary Applied Mathematics*. Springer New York, New York, NY, 2004.

[5] M. Hossein Mirabdollah and Bärbel Mertsching. Fast Techniques for Monocular Visual Odometry. In *German Conference on Pattern Recognition - DAGM 2015*, pages 297–307, 2015. 1, 3

[6] Raul Mur-Artal and Juan D. Tardos. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, oct 2017. 1

[7] Mikael Persson, Tommaso Piccini, Michael Felsberg, and Rudolf Mester. Robust stereo visual odometry from monocular techniques. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 686–691. IEEE, jun 2015. 1

[8] Shiyu Song and Manmohan Chandraker. Robust Scale Estimation in Real-Time Monocular SFM for Autonomous Driving. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1566–1573. IEEE, jun 2014. 1

[9] Nan Yang, Rui Wang, Jörg Stückler, and Daniel Cremers. Deep Virtual Stereo Odometry: Leveraging Deep Depth Prediction for Monocular Direct Sparse Odometry. In *European Conference on Computer Vision - ECCV 2018*, pages 835–852, 2018. 1