Relative 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

Determining the scale of relative motion is key to achieve consistency in monocular motion estimation when trajectories are recovered up to a scale factor. In this paper, we introduce a novel method to estimate the relative scale in monocular visual odometry using a calibrated camera. Our algorithm exploits redundancy in point depth information to achieve robust relative scale estimates. The performance of the method is evaluated in the KITTI public dataset for autonomous vehicles using the standard KITTI benchmark metrics. The results demonstrate the effectiveness of a robust relative scale estimation with 3.06% less drift against visual odometry without any scale correction, and a total average translation error of 33.23%.

1. Introduction

Visual odometry (VO) is the process of recovering the instantaneous position and orientation of a camera from sequences of images taken at successive instants [12]. In monocular visual odometry, a camera is used to estimate motion and without any knowledge of a reference metric of the scene under observation, the motion can only be determined up to a scale, which is known in literature as *relative scale estimation*. Estimating the relative scale is instrumental towards full autonomy in motion estimation because estimates of the relative scale can be combined with other sources of metric information to recover the absolute scale, for instance, using the camera height, the size of known objects in scene [13] or inertial sensors such as accelerometers and gyroscopes [10]. The specialized literature can be organized in two main groups: *Direct methods* and *Sensor fusion methods*. The direct methods use the information of a camera, while the sensor fusion methods merge visual content with other sources (e.g. stereo camera, inertial data, spatial metrics, non-holonomic constraints). We will limit our analysis to direct methods that aim to exploit purely monocular information.

Regarding direct methods, in [7] it was derived a batch estimation method that requires a set of m image frames with n 2–D points and the depth of the first point is used as reference for determining the relative scale. In [4], the authors used the *trifocal tensor* determined by groups of three related images. This method is analogous to the work of [7] for the case $m = 3$ images, however, it only requires point correspondences from three views. In [12], the authors expressed the relative scale as the mean of the distance ratio between pairs of triangulated points expressed in 3–D coordinates. Although several relative scale estimation mechanisms have been proposed in literature, such methods do not make use of redundant information given in the keypoint depths (direct methods), or rather use external information with sophisticated formulations (sensor fusion methods).

In this work, we provide an analytical solution for relative scale estimation based on a novel three-view triangulation algorithm. The proposed methodology uses a calibrated camera to derive a robust method for determining the relative scale by exploiting repeated observations of a 1–D parameter (the keypoint depth) rather than using full 3-D keypoint coordinates as in previous literature. The proposed algorithm is modular and can operate as a building block in other perception-oriented methodologies (e.g. visual SLAM). The performance of the algorithm is demonstrated on the KITTI dataset for autonomous vehicles [2].

2. Methodology

For a given point p in the 3-D space, we denote its representation in the world reference frame w by $\boldsymbol{X}^w = [x^w, y^w, z^w]^T$, in the camera reference frame c by $\boldsymbol{X}^c = [x^c, y^c, z^c]^T$, its perspective projection in normalized camera coordinates by $\boldsymbol{x} = [x^c/z^c, y^c/z^c, 1]^T$ and in image coordinates by $\hat{\boldsymbol{x}} = [u, v, 1]^T$. The relation of a 3-D point in the world with its 2-D projection on the image plane using the pinhole camera model in homogeneous coordinates

is given by [14]

$$z^c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} S_x f & 0 & u_0 \\ 0 & S_y f & v_0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathcal{K}} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathcal{R}^{cw} & \boldsymbol{t}^{cw} \\ \mathbf{0}^T & 1 \end{bmatrix}}_{[\mathcal{R}^{cw}|\boldsymbol{t}^{cw}]} \begin{bmatrix} x^w \\ y^w \\ z^w \\ 1 \end{bmatrix}, \tag{1}$$

where $\mathcal{K}$ is the camera calibration matrix, $\mathcal{R}^{cw} \in \mathbb{R}^{3\times3}$ is the rotation from world to camera coordinates, and $\boldsymbol{t}^{cw} \in \mathbb{R}^3$ is the translation of the world origin with respect to the camera. Hence, monocular visual odometry obeys the relation

$$\hat{\boldsymbol{x}} \sim \mathcal{C}\boldsymbol{X}^w, \tag{2}$$

where $\mathcal{C} = \mathcal{K}[\mathcal{R}^{cw}|\boldsymbol{t}^{cw}]$ is the camera projection matrix and $\sim$ indicates equality up to a scale factor given by the keypoint depth $z^c \in \mathbb{R}_+$.

A point p has coordinates $\boldsymbol{X}^c_{k-1}$ and $\boldsymbol{X}^c_k$ relative to a pair of consecutive camera frames, which are related by the rigid-body transformation

$$\boldsymbol{X}^c_k = \mathcal{R}_k \boldsymbol{X}^c_{k-1} + \boldsymbol{t}_k, \tag{3}$$

where the rotation $\mathcal{R}_k$ and translation $\boldsymbol{t}_k$ are the extrinsic parameters of camera motion. Those parameters can be estimated using well-established methods [7, 9]. However, the estimated motion may lead to inconsistent relative scales between frames since the keypoints involved in the calculations might be different. Hence, the computing the relative scale of the translation vector allows for consistent motion estimates when new views are incorporated into the scene.

2.1. Relative scale estimation

Fig. 1 depicts our procedure to estimate the relative scale. The relative scale of translation at frame instant k, denoted $\lambda^\star_k$, is estimated as the ratio of the depths z^c of a single 3-D point $\boldsymbol{X}^c_{k-1} = [x^c_{k-1}, y^c_{k-1}, z^c_{k-1}]^T$ and $\boldsymbol{X}^{c'}_{k-1} = [x^{c'}_{k-1}, y^{c'}_{k-1}, z^{c'}_{k-1}]^T$ triangulated at instant $k-1$ from consecutive pairs of camera views. Firstly, the relative motion from three camera views is recovered by decomposition of the essential matrices relative to a pair of consecutive views. Specifically, relative motion parameters $(\mathcal{R}_{k-1}, \hat{\boldsymbol{t}}_{k-1})$ are recovered from the frames $\{k-2, k-1\}$ by decomposition of the essential matrix $\mathcal{E}_{k-1}$ and $(\mathcal{R}_k, \hat{\boldsymbol{t}}_k)$ are recovered from the frames $\{k-1, k\}$ by decomposition of the essential matrix $\mathcal{E}_k$.

Secondly, the depths of the 3-D points $\boldsymbol{X}^c_{k-1}$ and $\boldsymbol{X}^{c'}_{k-1}$ are calculated by triangulation using the parameters of relative motion calculated previously. Particularly, depth z^c_{k-1}, corresponding to the former point, is recovered by triangulation using the frames $\{k-2, k-1\}$ and the depth $z^{c'}_{k-1}$, corresponding to latter point, is recovered by triangulation

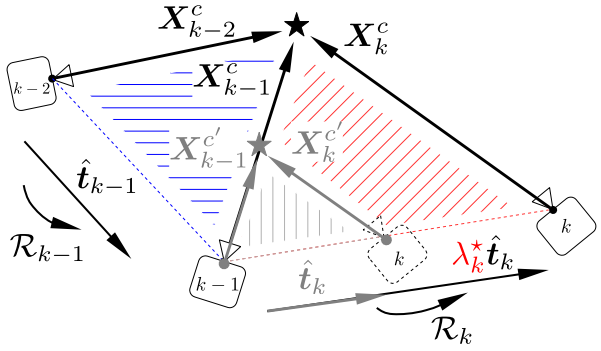

Figure 1: Relative scale estimation in monocular odometry. The striped lines in blue represent triangulation at instants $\{k-2, k-1\}$. The striped lines in gray represent triangulation at instants $\{k-1, k\}$. The relative scale $\lambda^\star_k$ is given by the ratio of depths from points $\boldsymbol{X}^{c'}_{k-1}$ and $\boldsymbol{X}^c_{k-1}$. The magnitude of the translation vector at the current frame k is scaled according to $\lambda^\star_k \boldsymbol{t}$ and represented by the striped lines in red.

using the frames $\{k-1, k\}$ [5]

$$\begin{bmatrix} z^c_{k-2} \\ z^c_{k-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}^T_{k-2}\boldsymbol{x}_{k-2} & -\boldsymbol{x}^T_{k-2}\mathcal{R}_{k-1}\boldsymbol{x}_{k-1} \\ -\boldsymbol{x}^T_{k-2}\mathcal{R}_{k-1}\boldsymbol{x}_{k-1} & \boldsymbol{x}^T_{k-1}\boldsymbol{x}_{k-1} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}^T_{k-2}\hat{\boldsymbol{t}}_{k-1} \\ -\boldsymbol{x}^T_{k-1}\mathcal{R}_{k-1}\hat{\boldsymbol{t}}_{k-1} \end{bmatrix} \tag{4a}$$

$$\begin{bmatrix} z^{c'}_{k-1} \\ z^c_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}^T_{k-1}\boldsymbol{x}_{k-1} & -\boldsymbol{x}^T_{k-1}\mathcal{R}_k\boldsymbol{x}_k \\ -\boldsymbol{x}^T_{k-1}\mathcal{R}_k\boldsymbol{x}_k & \boldsymbol{x}^T_k\boldsymbol{x}_k \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}^T_{k-1}\hat{\boldsymbol{t}}_k \\ -\boldsymbol{x}^T_k\mathcal{R}_k\hat{\boldsymbol{t}}_k \end{bmatrix}. \tag{4b}$$

A key step in our method is the fact that vectors $\boldsymbol{X}^c_{k-1}$ and $\boldsymbol{X}^{c'}_{k-1}$ refer to the same point at instant $k-1$ and, therefore, the estimated depths z^c_{k-1} and $z^{c'}_{k-1}$ are equal up to a relative scale factor since different keypoints might be involved in their computations. Thirdly, the estimated relative scale λ_k is given by the ratio of depths at frame $k-1$

$$\lambda_k = \frac{z^c_{k-1}}{z^{c'}_{k-1}}. \tag{5}$$

By exploiting redundancy in depth information, we derived an *optimal relative scale*, denoted $\lambda^\star_k$, as a least squares solution of (5) for three-view scenes with n corresponding keypoints, where $n \in \mathbb{N}$

$$\lambda^\star_k = \frac{\boldsymbol{Z}'^T_{k-1}\boldsymbol{Z}_{k-1}}{\|\boldsymbol{Z}'_{k-1}\|^2}, \tag{6}$$

where $\boldsymbol{Z}'_{k-1}$ and $\boldsymbol{Z}_{k-1}$ are vectors in $\mathbb{R}^n$ of the estimated keypoint depths, defined by

$$\boldsymbol{Z}'_{k-1} = [z^{c'}_{k-1,1}, z^{c'}_{k-1,2}, \ldots, z^{c'}_{k-1,n}]^T, \tag{7a}$$

$$\boldsymbol{Z}_{k-1} = [z^c_{k-1,1}, z^c_{k-1,2}, \ldots, z^c_{k-1,n}]^T. \tag{7b}$$

Finally, the translation vector is scaled up accordingly

$$\hat{\boldsymbol{t}}_k = \lambda_k^{\star} \hat{\boldsymbol{t}}_k \,. \tag{8}$$

2.2. Experimental setup

The performance of our method is evaluated on the KITTI dataset. The vehicle navigates in an outdoors environment characterized by a mostly static scene with the presence of natural features of an urban landscape, e.g. trees, cars, buildings. The coordinates of the reconstructed trajectory are expressed in a global reference frame with the origin centered in first camera frame. *Keypoint extraction* and *keypoint tracking* are implemented using the FAST descriptor [11] and the Lucas-Kanade tracker (LK-tracker) [6]. The extrinsic parameters are calculated using the 5-point algorithm [9] and the *3-D reconstruction* is calculated by triangulation as in [5].

3. Results

The evaluation of our algorithm in the standard KITTI dataset is shown in Table 1. The translational and rotational errors are calculated using the KITTI benchmarks, which are based on the average of different trajectory lengths at (100m, 200m, ..., 800m). The errors are measured in percent (%) for translation ($\boldsymbol{t}$) and in degrees per meter (deg /m) for rotation ($\mathcal{R}$). The total average error was 33.23 % for translation and 0.0006 deg /m for rotation. The results showed accurate scale estimates in scenarios with initial forward motion of the vehicle (sequences 00, 02, 03, 04, 05, 06, and 08). The algorithm is more susceptible when the initial motion corresponds to more challenging maneuvers as in turns (sequences starting with an initial rotation, e.g. 01, 07, 09, and 10). This may be associated with a faster loose of keypoint tracking during fast rotations in the initialization with a corresponding affectation on the estimated motion parameters and, consequently, poor relative scale estimates.

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

ID	**L** (m)	**Environment**	$\boldsymbol{t}$ (%)	$\mathcal{R}$ (deg /m)
00	3714	Urban	17.03	0.0007
01	4268	Highway	64.15	0.0004
02	5075	Urban+Country	35.92	0.0006
03	563	Country	28.99	0.0003
04	397	Country	5.62	0.0001
05	2223	Urban	18.97	0.0004
06	1239	Urban	7.75	0.0002
07	695	Urban	43.86	0.0006
08	3225	Urban+Country	19.39	0.0006
09	1717	Urban+Country	54.00	0.0005
10	919	Urban+Country	69.86	0.0008

4. Analysis

The family of batch methods, which are based on [7], use m-view reconstruction, which may lead to poor relative scales (estimates depend on the number of matched points and such number decreases with $m > 2$). In contrast, we proposed a sequential estimation from pairs of camera frames (2-view reconstruction) and such approach imposes only two bilinear constraints per iteration aiming for increased robustness against data noise.

Compared to the more similar family of online methods, which are based on [12], both algorithms are asymptotically equal since they have linear complexity. However, for a sufficiently large number of points, our solution requires 20 times less operations per iteration (1-D depths against full 3-D vectors) and such result might be relevant in platforms with limited computational resources for online processing.

Our algorithm achieved 3.06% less drift against visual odometry without any scale correction and 33.23% on average translational error in the KITTI ranking against state-of-the art results of up to 21.47% [3]. Unlike current monocular VO approaches reported in [3], our proposal is independent of hard prerequisites, such as external information (machine learning), additional sensors (stereo vision, sensor fusion), offline processing (pose-refinement, batch-processing), or any combination of them. This may explain the difference in accuracy, but in contrast, our method can be easily incorporated in any other monocular VO approach to further improve the estimated odometry.

The performance of our relative scale estimation algorithm on 1 core@2.3 GHz (Python) was 0.2453 seconds per iteration. Considering this, we argue that the average runtime is within the top 10 rank of the fastest methods [3].

5. Conclusion

We proposed a novel method for estimating the relative scale in monocular visual odometry using a calibrated camera. By introducing robust relative scales, we achieved near state-of-the art results in the KITTI ranking (33.23% of average translation error). The robust relative scale estimates were obtained by exploiting redundancy in depth information of points in the scene, which was shown to be key in reducing the odometry drift. Our method can be implemented in other related perception-oriented approaches, such as the visual SLAM, and is feasible of real-time implementation.

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.

[2] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

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

[4] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, 2004. 1

[5] Paulo Roberto Gardel Kurka and Aldo André Díaz Salazar. Applications of image processing in robotics and instrumentation. *Mechanical Systems and Signal Processing*, 124:142–169, jun 2019. 2, 3

[6] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, pages 121–130, 1981. 3

[7] 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. 1, 2, 3

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

[9] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, jun 2004. 2, 3

[10] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, aug 2018. 1

[11] Edward Rosten, Reid Porter, and Tom Drummond. Faster and Better: A Machine Learning Approach to Corner Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):105–119, jan 2010. 3

[12] Davide Scaramuzza and Friedrich Fraundorfer. Visual Odometry [Tutorial]. *IEEE Robotics & Automation Magazine*, 18(4):80–92, dec 2011. 1, 3

[13] 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

[14] Richard Szeliski. *Computer Vision: Algorithms and Applications*, volume 5 of *Texts in Computer Science*. Springer, New York, NY, 2011. 2