Euclidean and Riemannian Metrics in Learning-based Visual Odometry

Olaya Álvarez-Tuñón, Yury Brodskiy, and Erdal Kayacan

***Abstract*—This paper overviews different pose representations and metric functions in visual odometry (VO) networks. The performance of VO networks heavily relies on how their architecture encodes the information. The choice of pose representation and loss function significantly impacts network convergence and generalization. We investigate these factors in the VO network DeepVO by implementing loss functions based on Euler, quaternion, and chordal distance and analyzing their influence on performance. The results of this study provide insights into how loss functions affect the designing of efficient and accurate VO networks for camera motion estimation.**

***Index Terms*—Visual Odometry, Deep Learning, Lie algebra, Riemannian geometry**

I. Introduction

Several deep learning architectures have arisen in the last decade to support learning from spatio-temporal data, which have allowed the implementation of architectures for end-to-end visual odometry (VO) methods [1]. PoseNet [2] introduced end-to-end visual localization with a convolutional neural network (CNN) pretrained for classification. Later architectures introduced optical flow networks followed by CNN pose regressors [3], [4], long short-term memory (LSTM) units [3] and self-attention mechanisms [5].

The choice of loss function significantly impacts the information encoded by the network. In VO, the main challenge to address when implementing loss functions is the rotation representation [6]. The supervised pose regression networks usually rely on the Euclidean loss for rotation and translation [2], [3], [5]. These methods implement different variations of a weighted sum of the Euclidean loss for the rotation and translation components. PoseNet [2] and Atloc [5] represent rotations with quaternions and DeepVO [3] outputs Euler angles.

Using the state-of-the-art VO network DeepVO as backbone, this research examines various pose representations and their corresponding loss functions, as well as the impact they have on the network's convergence and generalization ability.

II. Geometry for visual odometry

The geometry representation that best models the perspective projection from camera imaging is projective geometry. Projective transforms only preserve type (points or lines), incidence (whether a point lies in a line) and cross-ratio. Euclidean geometry is a subset of projective geometry that, in addition to that, also preserve lengths, angles and parallelism. There are two other hierarchies between them: affine and similarity geometry, as shown in Fig. 1. Projective reconstruction leads to distorted models; thus visual localization algorithms aim to estimate the Euclidean structure of the scene [7]. Moreover, VO requires quantifying differences in translation and orientation between camera frames.

A. Pose parameterization and optimization

Pose representation involves translation and rotation with respect to a reference frame, which can be parameterized using Euler angles, quaternions, or rotation matrices. Euler angles are intuitive but can lead to numerical instabilities and non-linearities, making optimization difficult. Quaternion is a better representation when compared to Euler angles which avoids gimbal lock problems and is more numerically stable. Unitary quaternions belong to the special unitary group $SU(2)$. However, quaternions double-cover the space of rotations. Rotation matrices belong to the special orthogonal group $SO(3)$, and their combination with position information forms the special Euclidean group $SE(3)$. Unlike the quaternions, the elements in the $SO(3)$ group uniquely represent rotations in the space. Rotation and transformation matrices can be considered elements of a Lie group, forming a smooth manifold that can be smoothly parameterized.

B. Distances in the Euclidean space

Learning-based VO requires finding an appropriate metric that can accurately and robustly quantify the quality of the estimated camera motion. In mathematics, a metric or distance $d(A, B)$ is a way to measure the distance between two points on any set. Furthermore, it must satisfy non-negativity,

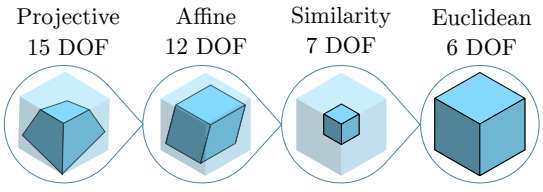

Fig. 1. Hierarchy of geometries. Euclidean preserves areas, angles and lengths, the similarity preserves ratios of lengths. The affine preserves volumetric ratios and parallelism. The projective preserves intersections and tangents.

O. Álvarez-Tuñón is with Artificial Intelligence in Robotics Laboratory (AiRLab), the Department of Electrical and Computer Engineering, Aarhus University, 8000 Aarhus C, Denmark {olaya at ece.au.dk}. Y. Brodskiy is with EIVA a/s, 8660 Skanderborg, Denmark {ybr at eiva.com}. E. Kayacan is with Automatic Control Group (RAT), Paderborn University, 33098 Paderborn, Germany {erdal.kayacan at uni-paderborn.de}.

identity, symmetry, and triangle inequality properties. In differential geometry, it is specifically defined on a geometric space, e.g., a Euclidean space or a Riemannian manifold.

Translations have a straightforward geometric interpretation as displacements in space. They can be measured by satisfying the properties of a metric by the Euclidean distance. On the other hand, rotations do not have a direct geometric interpretation and require other mechanisms to find a metric. Therefore, the following section discusses prevalent metrics and pseudo-metrics used in the literature for the proposed parameterizations for orientations.

C. Distances in the vector space of orientations

Some well-known distances used for 3D rotations in the VO and SLAM literature are the Euclidean distance, the geodesic distance and the cordal distance, applied to the different orientation parameterizations as follows [8]:

1) Distance on Euler angles: two three-dimensional vectors of Euler angles φ_A and φ_B yield the Euclidean distance as:

$$d_e(\varphi_A, \varphi_B) = ||\varphi_A - \varphi_B||_2^2 \tag{1}$$

with $||\cdot||_2$ the L_2 norm. However, it cannot be considered a metric as it does not satisfy the triangle inequality. This induces singularities during the optimization, potentially leading to suboptimal solutions or slow convergence.

2) Quaternion distances:

a) Euclidean distance: Two unit quaternions q_A and q_B yield the Euclidean distance:

$$d_q(q_A, q_B) = \|q_A - q_B\|_2^2 \tag{2}$$

Due to the quaternion's double-cover, q_B and $-q_B$ represent the same rotation but do not retrieve the same distance, i.e.: $d(q_A, q_B) \neq d(q_A, -q_B)$. This matter can be addressed by redefining the quaternion distance as:

$$d_q e(q_A, q_B) = min_{b\in\{-1;+1\}}\|q_A - bq_B\|_2^2 \tag{3}$$

which defines a pseudo-metric since it does not satisfy the identity property. Furthermore, it incorporates the problem of adding a binary variable to the equation, hindering the computational analysis.

b) Geodesic distance: The Riemannian metric on $SU(2)$ induces a distance on the quaternion manifold obtained as the geodesic distance:

$$d_q g(q_A, q_B) = ||log(q_A^{-1} q_B)||_2^2 \tag{4}$$

3) Distances in Lie groups:

a) Geodesic distance: The distance between two rotations $R_A \in SO(3)$ and $R_B \in SO(3)$ can be obtained as the rotation angle θ_{AB} corresponding to the relative rotation $R_{AB} = R_A^T R_B$:

$$d_\theta(R_A, R_B) = \|arccos\left(\frac{tr\left(R_A^T R_B\right) - 1}{2}\right)\| \tag{5}$$

The norm of the exponential coordinates is the rotation angle; thus, the previous metric can be written as:

$$d_\theta(R_A, R_B) = \|log(R_A^T R_B)^\vee\| = \|log(R_B^T R_A)^\vee\| \tag{6}$$

This distance is geodesic, i.e., the length of the minimum path between R_A and R_B on the $SO(3)$ manifold. The geodesic distance defines a Riemannian metric that satisfies the metric properties. Moreover, it defines a smooth metric since both the logarithmic map and the Euclidean norm are smooth. However, it brings more computational expense and numerical instability from the logarithm map for a set of big rotations.

b) Chordal distance: The chordal distance between two rotations $R_A \in SO(3)$ and $R_B \in SO(3)$ is defined as [6]:

$$d_c(R_A, R_B) = \|R_A - R_B\|_F = \|R_A R_B^T - I\|_F \tag{7}$$

with $\|\cdot\|_F$ the Frobenius norm. In most applications, minimising the square of the chordal distance is preferred. The chordal distance is not a Riemannian metric, but it complies with the four metric requirements while being more numerically stable and simpler than the geodesic distance [9].

III. Experiments

Considering the distances introduced above, this section proposes a series of experiments to compare them.

A. Experiment setup

The network chosen to investigate the influence of the pose loss functions in the model's performance is DeepVO [3]. DeepVO yields a pre-trained FlowNet [10] CNN followed by two LSTM units. DeepVO is well suited for the proposed experiment since it only requires a loss purely dependent on the distance between the target and the estimated pose.

Four experiments are carried out, consisting in training DeepVO under the same training setup for three different orientation parameterizations: Euclidean angles, quaternions, and SO(3) yielding the SE(3) pose.

The three different parameterizations are used to train four models with four different loss functions. The loss function proposed in DeepVO corresponds to the mean squared error (MSE) of the Euclidean loss, with the orientation represented as Euler angles:

$$L_{original} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}||t_i - \hat{t}_i||_2^2 + k_1||\varphi_i - \hat{\varphi}_i||_2^2 \tag{8}$$

where for each observation i, $\hat{t}_i$ and $\hat{\varphi}_i$ are the estimated values for rotation and orientation with ground truth t_i and φ_i. M and N correspond to the sequence length and the number of observations, respectively.

The Euclidean loss for the pose represented as translation and quaternion vectors is obtained as:

$$L_{quat\_e} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}||t_i - \hat{t}_i||_2^2 + \min_{b\in\{-1;+1\}} k_2||q_i - b\hat{q}_i||_2^2 \tag{9}$$

with q_i and $\hat{q}_i$ the ground truth and estimated quaternions. For this experiment, the output head of the network is modified to provide seven outputs corresponding to the three translation

elements and the four rotation elements. The geodesic loss for the quaternion parameterization yields:

$$L_{quat\_geo} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} ||t_i - \hat{t}_i||_2^2 + k_3||log(q_A^{-1} q_B)||_2^2 \quad (10)$$

Finally, the chordal loss for the SE(3) representation is obtained as:

$$L_{SE(3)} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} \|t_i - \hat{t}_i\|_2^2 + k_4 d_c(R_i, \hat{R}_i)^2 \quad (11)$$

with R_i and $\hat{R}_i$ the ground truth and estimated SO(3) rotations. It is to be noted that the rotation matrix is obtained from the network's output vector through the exponential map.

The three constants k_1,k_2, k_3, and k_4 are experimentally determined. They aim to equate the order of magnitude of the orientations to that of the translations. For these experiments, $k_1 = 100$, $k_2 = 14$, $k_3 = 100$, and $k_4 = 153$.

The training is carried out with the KITTI VO dataset [11]. We select the sequences 00, 02-08, and 08 as training set and the sequences 09 and 07 as validation set. With a pre-trained FlowNet as initial checkpoint, the network is trained for up to 200 epochs with a 20% dropout and using an Adam optimizer with a learning rate of 0.001, as originally proposed by DeepVO [3].

B. Results and discussion

The train and validation losses for the three experiments are shown in Fig. 2. First, the graphic illustrates a lack of data for adequate convergence. Due to time limitations, feeding more data to the network and tuning the regularization parameters are left as future works. Instead, these experiments are taken as a pure comparison of the losses' performance under the same conditions. Figure 2 shows a better convergence of the loss function under the chordal loss from the SE(3) parameterization, evidenced by a steady decrease over the 200 epochs. This decrease is present in training and validation, which indicates better generalization. As opposed to that, the loss under the Euler angle parameterization presents a rapid but mild decrease of the loss for the first epochs during training, to then enter a flat area. The loss under the quaternion parameterization shows a very slow convergence where it seems to converge to a local minimum between epochs 25 and 100, to then steadily decrease again.

The trajectory plots in Fig 3 show the superiority of the models trained with quaternions and SE(3) representation versus the original implementation. The Euler angles representation presents big rotation shifts even in the data seen during training. The model using quaternions and SE(3) representation shows similar performance. However, the model using SE(3) shows a better resemblance of the orientations in the trajectories that becomes more noticeable in Trajectory 09. We observe big orientation drifts under steep rotations on the Trajectories 03 and 10. This is due to the low presence of steep rotations in the dataset, which hinders learning such data distributions. In conclusion, the SE(3) representation with a chordal loss presents the best performance under the fastest convergence.

IV. Conclusions

This study demonstrates that using the chordal loss under SE(3) pose representation provides better convergence and generalization compared to the Euclidean loss under quaternion and Euler angles parameterization. This highlights the strong influence of the loss function on the information encoded by the network, and how the choice of loss function conditions its ability to converge and generalize. Future works will involve adding more data, performing hyperparameter optimization, and implementing more losses in the Riemannian and Euclidean space.

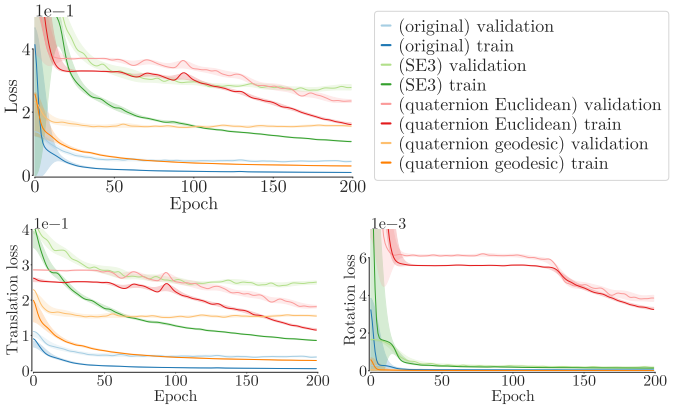

Fig. 2. Top: pose loss values for the pose loss during train and validation. Bottom: Translation and rotation losses (without weighting). Note that the rotation losses do not have the same geometric interpretation.

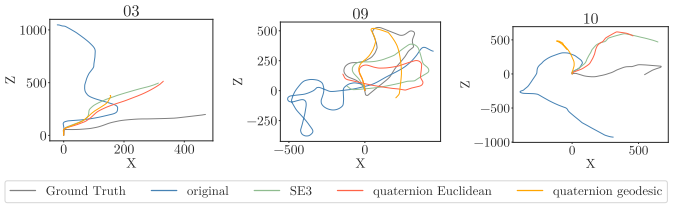

Fig. 3. From left to right, trajectories used for training, validation, and test. The plot shows the ground truth versus estimated pose for DeepVO using the proposed pose representations and loss functions.

Acknowledgment

This paper is written under the project REMARO which has received funding from the European Union's EU Framework Programme for Research and Innovation Horizon 2020 under Grant Agreement No 956200.

References

[1] P. Du, X. Bai, K. Tan, Z. Xue, A. Samat, J. Xia, E. Li, H. Su, and W. Liu, "Advances of four machine learning methods for spatial data handling: A review," *Journal of Geovisualization and Spatial Analysis*, vol. 4, no. 1, pp. 1–25, 2020.

[2] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[3] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 2043–2050.

[4] W. Wang, Y. Hu, and S. Scherer, "Tartanvo: A generalizable learning-based vo," *arXiv preprint arXiv:2011.00359*, 2020.

[5] B. Wang, C. Chen, C. X. Lu, P. Zhao, N. Trigoni, and A. Markham, "Atloc: Attention guided camera localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 06, 2020, pp. 10 393–10 401.

[6] L. Carlone and F. Dellaert, "Duality-based verification techniques for 2d slam," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 4589–4596.

[7] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[8] L. Carlone, R. Tron, K. Daniilidis, and F. Dellaert, "Initialization techniques for 3d slam: A survey on rotation estimation and its use in pose graph optimization," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 4597–4604.

[9] I. Aloise and G. Grisetti, "Matrix difference in pose-graph optimization," *arXiv preprint arXiv:1809.00952*, 2018.

[10] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.