Fast and Robust ℓ_1-averaging-based Pose Estimation for Driving Scenarios

German Ros[1,2]
gros@cvc.uab.es

Julio Guerrero[3]
juguerre@um.es

Angel D. Sappa[1]
asappa@cvc.uab.es

Daniel Ponsa[1,2]
daniel@cvc.uab.es

Antonio M. López-Peña[1,2]
antonio@cvc.uab.es

[1] Computer Vision Center
Edifici O, Campus UAB, 08193
Bellaterra (Barcelona), Spain

[2] Computer Science Dept.
Universitat Autònoma de Barcelona
Campus UAB, Bellaterra (Barcelona), Spain

[3] Dept. de Matemática Aplicada
FIUM, Universidad de Murcia
Campus de Espinardo, Murcia, Spain

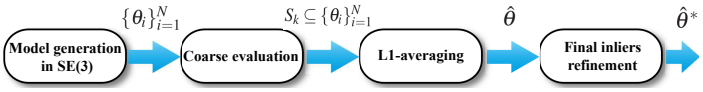

Figure 1: Main pipeline stages of the presented technique.

Robust camera-pose estimation is a fundamental stage of many computer vision problems, being specially important for Visual Simultaneous Localization and Mapping (*VSLAM*) [1][6] and Visual Odometry (*VO*)[4][5] systems. Here, the capability of estimating correct models in the presence of noise and high level of outliers is a fundamental requirement. During the last decades, many approaches have been proposed to solve these problems, being RANSAC [2] one of the most accepted and used. However, with the arrival of new challenges, such as large driving scenarios for autonomous vehicles, along with the improvements in the data gathering frameworks, new issues must be considered. One of these issues is the capability of a technique to deal with very large amounts of data while meeting the real-time constraint.

The use of large amounts of data to perform model estimation has proven to be beneficial for improving model accuracy, as stated in [7]. In the past, the amount of input information was very limited, but now modern front-ends are able to extract thousands of features from a set of images and match them in real-time in a standard CPU. Nevertheless, serious issues, as the presence of outliers within the flow of data still need to be carefully addressed. RANSAC-like methods are affected by this "excess" of information, what produces an increment on the time dedicated to evaluate and rank the generated models. In order to avoid this drawback, real-time implementations opt to use just a part of the available data, therefore discarding a great amount of information and penalizing the accuracy of the resultant models. To address these problems, in the presented work we propose a novel technique to perform robust camera-pose estimation that is specially suitable to deal with large flows of data, something that is common in urban scenarios, where we successfully tested the approach.

The main idea behind our proposal is to combine a very fast hypotheses assessment that produces sub-optimal evaluation and a procedure capable of combining partially incorrect hypotheses into a new and accurate hypothesis. The procedure is depicted by Fig. 1, and starts by creating N candidate models $\theta_i \in \mathbb{SE}(3)$ from the available data $\mathcal{X} = \{(x_{l,p}, x_{r,p}, x_{l,c}, x_{r,c})^{(i)}\}_{i=1}^D$, which represent matched points in the four views of a moving stereo-rig (i.e., at two different time instants). This step is carried out by optimizing Eq. 1 for $M = 3$ correspondences while accounting for the manifold structure of the model.

$$\text{argmin}_\psi \Sigma_{i=1}^M \left\| \text{K}\Pi_3 \left(\mathbf{exp}_r(\psi) \hat{X}_{l,p}^{(i)} \right) \times \hat{x}_{l,c}^{(i)} \right\|_{\ell_2}^2 + \left\| \text{K} \left(\Pi_3 \left(\mathbf{exp}_r(\psi) \hat{X}_{l,p}^{(i)} \right) - \vec{B} \right) \times \hat{x}_{r,c}^{(i)} \right\|_{\ell_2}^2 \quad (1)$$

After generating the set of models $\{\theta_i\}_{i=1}^N$, they are assessed by F_{coarse} (Eq. 2), an evaluation function that has been designed to be extremely fast. This is achieved thanks to the use of a *Reduced Measurement Matrix* (*RMM*) [3], an algebraical reduction of the input data $\mathcal{X}$ that creates a compact equivalent $\boldsymbol{M}$ under the ℓ_2-norm. The advantage of this reduction is that $\boldsymbol{M}$ can be efficiently computed even for very large collections of data and this has to be done just once, at the beginning of the process.

$$F_{coarse}(\theta) = \Sigma_{i=1}^D \left\| \text{K} \left(\Pi_3 \theta \hat{X}_{l,p}^{(i)} \right) \times \hat{x}_{l,c}^{(i)} \right\|_{\ell_2}^2 + \left\| \text{K} \left(\Pi_3 \theta \hat{X}_{l,p}^{(i)} - \vec{B} \right) \times \hat{x}_{r,c}^{(i)} \right\|_{\ell_2}^2 =$$

$$\Sigma_{i=1}^D \left\| W_l^{(i)} \check{\theta} \right\|_{\ell_2}^2 + \left\| W_r^{(i)} \check{\theta} \right\|_{\ell_2}^2 = \left\| \boldsymbol{W_l} \check{\theta} \right\|_{\ell_2}^2 + \left\| \boldsymbol{W_r} \check{\theta} \right\|_{\ell_2}^2 = \check{\theta}^T (\boldsymbol{M_l} + \boldsymbol{M_r}) \check{\theta} \quad (2)$$

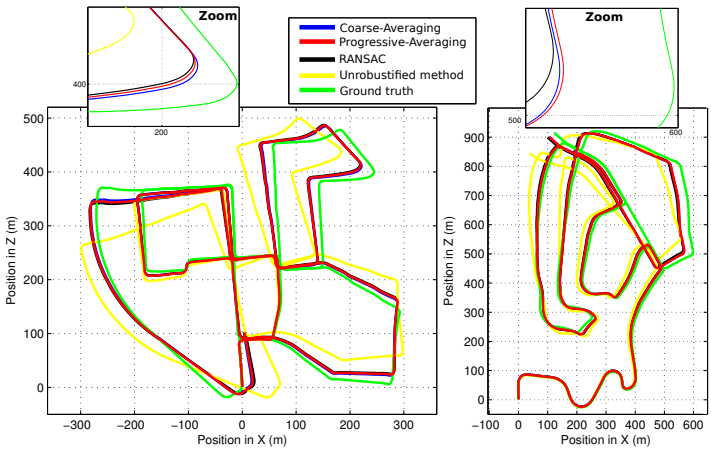

Figure 2: Results of the *VO* experiments for KITTI sequences 00 (left) and 02 (right). Notice that, although all robust methods lead to similar trajectories, C-Avg and P-Avg remain closer to the ground truth.

However, F_{coarse} is not robust to noise and outliers, what will lead to the wrong evaluation of some models. Fortunately, a thorough analysis of real data sequences showed that models with a high number of inliers (the expected good models) produce low residuals for F_{coarse}, even when $\boldsymbol{M}$ contains outliers. On the other hand, models corresponding with a low number of inliers present random values for F_{coarse}, producing low residuals just occasionally. Our approach exploits this weak property by selecting a subset $S_k \subseteq \{\theta_i\}_{i=1}^N$ of partially corrupted models according to the decision of F_{coarse}.

Afterwards, all the models are combined in a robust and fast way to generate a final and more accurate model $\hat{\theta}$. This is done by using the Weiszfeld algorithm to perform ℓ_1-averaging on $\mathbb{SE}(3)$ in order to respect the structure of the camera pose. We show that, under the appropriate conditions, the resultant model $\hat{\theta}$ is very accurate, at the level of models estimated by RANSAC, as shown in Fig. 2. The benefits of using this approach against RANSAC are shown in the paper.

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE TPAMI*, 2007.

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6), 1981.

[3] R. Hartley. Minimizing algebraic error in geometric estimation problems. In *Proceedings of the IEEE ICCV*, Washington, DC, USA, 1998.

[4] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In *Proceedings of the IEEE IV Symposium*, San Diego, CA, USA, 2010.

[5] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the IEEE CVPR*, volume 1, Washington, DC, USA, 2004.

[6] G. Sibley, C. Mei, I. Reid, and P. Newman. Vast-scale outdoor navigation using adaptive relative bundle adjustment. *International J. of Robotics Research*, 2010.

[7] H. Strasdat, J. M. M. Montiel, and A. J. Davison. Visual SLAM: Why filter? *Image Vision Comput.*, 30(2):65–77, February 2012.