Robust Real-Time Visual Odometry with a Single Camera and an IMU

Laurent Kneip
laurent.kneip@mavt.ethz.ch
Margarita Chli
margarita.chli@mavt.ethz.ch
Roland Siegwart
rsiegwart@ethz.ch

Autonomous Systems Lab
ETH Zurich, Switzerland

The increasing demand for real-time high-precision Visual Odometry (VO) systems as part of navigation and localization tasks has recently been driving research towards more versatile and scalable solutions. In this paper, we present a novel framework for combining the merits of inertial and visual data from a monocular camera to accumulate estimates of local motion incrementally and reliably reconstruct the trajectory traversed.

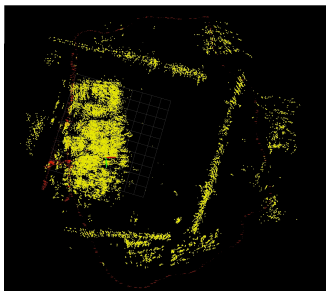

Figure 1: Map created by our VO implementation.

It has long been acknowledged that the use of inertial sensors together with cameras can complement each other in challenging scenarios, aiding the resolution of ambiguities in motion estimation arising when using each modality alone [6]. Here, we employ a monocular camera and an Inertial Measurement Unit (IMU) to recover relative camera motion, in a sensor setup available in practically most modern smart phones (*e.g.* iPhone, Google phones). We use a RANSAC based [1] hypothesize-and-test procedure as in [5], only here we specifically address the efficiency and robustness of monocular VO, presenting an elegant framework which exploits the additional benefits of the available rotation information. Figure 1 shows a map that has been incrementally generated by our VO implementation.

As shown in several recent works [2, 3, 4], knowledge about the vertical direction can for instance be used for reducing the minimum number of points for instantiating a hypothesis about the relative camera pose down to three or even only two in the perspective pose computation case. However, even though the vertical direction can be obtained from inertial data, it only works reasonably well in the static case. In this work, we propose an alternative tightly coupled SFM approach, that incorporates short-term full 3D relative rotation information from inertial data in order to support the geometric computation.

We follow a keyframe-based methodology such that whenever the median-filtered disparity exceeds a certain threshold in the number of pixels, we triangulate a new point cloud and the two frames used for this process serve as *keyframes*. In subsequent frames, the current pose is estimated with respect to the scene model constructed from the most recent keyframe-pair until a new keyframe is added to the system. Both the initial relative frame-to-frame transformation and the perspective pose computation are performed within robust RANSAC outlier-rejection schemes. The novel methodology presented here has been specifically designed to increase both the efficiency and the robustness of monocular VO estimation by exploiting priors about the relative rotations from an additional IMU. The major difference to [5] and also the key-contribution of this paper, is that this allows for a reduction in the number of points used in the RANSAC-hypotheses down to a minimum of two, for both the 2D-to-2D correspondence estimation during initialization and the 3D-to-2D correspondence problem upon perspective pose computation. These two-point algorithms do not suffer from any geometric degeneracies and always return a unique solution.

Knowledge of the relative camera rotation also provides great benefit during keyframe generation since disparities due to rotation can be compensated for, and as a result, the method can guarantee that there is enough translation between keyframe-pairs (boosting robustness of triangulation) despite that there is no prior information about the structure of the scene that the camera is exploring.

We demonstrate the successful application of our pose estimation methodology on data obtained using a Micro Aerial Vehicle (MAV) exhibiting full 3D motion. Compared to handheld camera motion sequences or images taken from a ground or fixed wing aerial vehicle, this setup provides very challenging datasets since the horizontal acceleration of the vehicle can directly be translated into roll and pitch rotations only. Since the methodology presented obtains relative rotation priors from the IMU, our framework is able to robustly cope with such critical motion sequences in contrast to classical vision-only based solutions. Moreover, the distribution of the extracted features in the image can be very inhomogeneous. Since the visual information is mainly used for estimating the translation, the algorithm is able to robustly work even with such uneven distributions.

The performance of the proposed system is assessed in terms of both speed and accuracy with respect to ground truth. Our results demonstrate minimal accumulated drift in estimates, presenting a relative assessment of different state-of-the-art feature types. The absolute error at the end of the 22.59m long trajectory amounts to 1.28m for the combination FAST+BRIEF (5.7%), 0.57m for AGAST+BRIEF (2.5%), 0.67m for AGAST+SURF (3%), and 0.26m for SURF+SURF (1.2%). As a result, we conclude that the choice of feature detector has the biggest impact on the overall success of operation. The ranking of combinations in terms of quality of results is evidently in contrast to the ranking in terms of computational efficiency. As shown in Figure 2, only the FAST or AGAST based solutions are able to run in real-time. The best trade-off between accuracy and efficiency is given by using the AGAST extractor with subpixel refinement in combination with the efficient BRIEF descriptor.

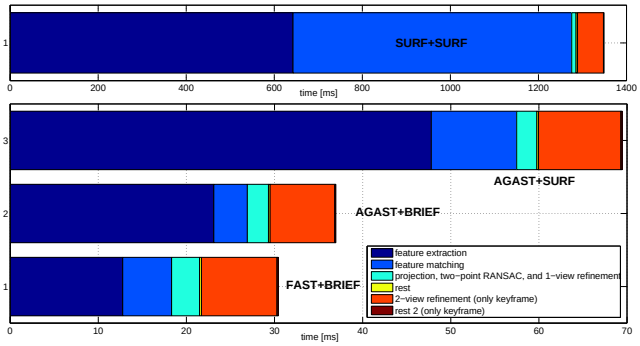

Figure 2: Timings of the different combinations of feature detectors and descriptors.

[1] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[2] F. Fraundorfer, P. Tanskanen, and M. Pollefeys. A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.

[3] M. Kalantari, A. Hashemi, and F. Jung J.-P. Guedon. A new solution to the relative orientation problem using only 3 points and the vertical direction. *Journal of Mathematical Imaging and Vision (JMIV)*, 39: 259–268, 2011.

[4] Z. Kukelova, M. Bujnak, and T. Pajdla. Closed-form solutions to the minimal absolute pose problems with known vertical direction. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2010.

[5] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[6] D. Strelow and S. Singh. Motion estimation from image and inertial measurements. *International Journal of Robotics Research (IJRR)*, 23(12):1157, 2004.