Robust Embedded Egomotion Estimation

Rainer Voigt, Janosch Nikolic, Christoph Hürzeler, Stephan Weiss, Laurent Kneip and Roland Siegwart

Abstract—This work presents a method for estimating the egomotion of an aerial vehicle in challenging industrial environments. It combines binocular visual and inertial cues in a tightly-coupled fashion and operates in real time on an embedded platform. An extended Kalman filter fuses measurements and makes motion estimation rely more on inertial data if visual feature constellation is degenerate. Errors in roll and pitch are bounded implicitly by the gravity vector. Inertial sensors are used for efficient outlier detection and enable operation in poorly and repetitively textured environments. We demonstrate robustness and accuracy in an industrial scenario as well as in general indoor environments. The former is accompanied by a detailed performance evaluation supported with ground truth measurements from an external tracking system.

Index Terms – Egomotion Estimation, Extended Kalman Filter, Stereo Vision, Vision-IMU Fusion, Visual Odometry

I. INTRODUCTION

The goal of the AIRobots project (www.airobots.eu) is the development of aerial vehicles capable of performing various inspection tasks in industrial environments. These include indoor setups of the power production industry such as very large boilers (up to 100 m high), cooling towers and flares. One of the main challenges in the employment of micro aerial vehicles for these tasks is accurate localization and egomotion estimation.

Performance requirements, payload limitations, blocked satellite reception and disturbances of the Earth’s magnetic field limit choices for the onboard sensor suite. External errors in roll and pitch are bounded implicitly by the gravity vector. Inertial sensors are used for efficient outlier detection and enable operation in poorly and repetitively textured environments. We demonstrate robustness and accuracy in an industrial scenario as well as in general indoor environments. The former is accompanied by a detailed performance evaluation supported with ground truth measurements from an external tracking system.

Fig. 1. Experimental sensor head including two global shutter CMOS imagers with 90° field of view and a custom low-cost MEMS based IMU.

To tackle these problems, this work suggests a robust method based on stereo vision and inertial measurements. Past research in the area of visual odometry has shown that vision and inertial measurements complement each other very well [6]. Attitude error can be bounded in two axes through the gravity vector [5], and visual estimates provide corrections for sensor drift. We propose a tight coupling between vision and IMU allowing us to robustly track our pose (position and attitude) in situations where purely vision based tracking fails. In the scenario of interest, the structure is generally close to the vehicle compared to the stereo baseline. This makes the visual pose estimation problem well constrained, and the absolute scale of structure and motion can be obtained immediately. This also holds for newly observed features, allowing for efficient outlier removal schemes and fast reinitialization after a tracking loss. Fig. 1 depicts the prototype sensor head used in our experiments.

II. VISUAL MOTION ESTIMATION

The core of the proposed odometry system is egomotion estimation using the stereo camera. The relative motion between consecutive stereo frames is computed, yielding
a chain of transformations that represent our current pose relative to the pose of the very first image pair. The algorithm consists of three steps.

A. Feature Detection and Triangulation

Initially, interest points are detected in the left and right image of a stereo frame. An enhanced version of the FAST [7] feature detector is employed, offering high repeatability at low computational cost. Instead of a single threshold for the entire image an adaptive thresholding scheme is proposed. The image is divided into a grid of square cells and the corner threshold of each cell is decreased individually until at least one feature per cell is detected. Adaptive thresholding comes at around twice the computational cost of the fixed threshold implementation as corners for each cell need to be extracted twice on average. However, spatially well distributed feature sets can be extracted in case of uneven image contrast. This yields good constraints for pose estimation.

In a next step computationally efficient BRIEF [8] descriptors are extracted for each feature. Using the epipolar constraint, left-to-right feature matching is performed. Only mutually best matches are considered valid features and are triangulated.

B. Frame to Frame Correspondence Search

Stereo matching and triangulation is performed for each successive stereo frame. Then, frame to frame correspondences are established by matching the descriptors of the current features to the descriptors from the previous frame. Again, only mutually best matches are accepted as valid correspondences. Furthermore, the frame to frame correspondence search is heavily constrained by motion estimates based on inertial measurements, see Fig. 3. This significantly reduces false matches and allows for reliable matching even in repetitive and poorly textured environments. The techniques used are outlined in V.

C. Pose Estimation

A RANSAC [9] approach is used to find an inlier triple of frame to frame feature correspondences. Based on the randomly drawn correspondence triples, camera motion is computed via a fast P3P algorithm [10] which generally yields four solutions for the motion. Two solutions with opposite camera orientation are discarded immediately. For the remaining two poses, all features are projected into the current frame. The score for the pose is computed from the number of supporting correspondences. As the pure correspondence count is not a good quality measure for the motion hypothesis, density filtering is applied in the image plane. Features that are close together are counted only once and the score function therefore also reflects the quality of the feature distribution in the images. The density filtering method ensures that we prefer to use hypotheses with fewer, evenly spread inliers over hypotheses with large numbers of inliers concentrated on a small, highly textured surface.

The pose with the best score from the RANSAC step is finally refined using bundle adjustment [11] over the current and previous frame. Bundle adjustment is performed iteratively in order to further remove outliers. Each time the adjustment has converged we check if there are any constraints violating the desired reprojection error bounds. If so, the feature causing the largest reprojection error is removed from the inlier set and optimization continues. If after convergence all constraints fulfill the error bound, iteration stops and the current solution is accepted.

III. Vision System Covariance Recovery

For robust motion estimation it is desirable to not only obtain a point estimate of the camera motion, but also an estimate of the uncertainty associated with this solution. Knowledge of this uncertainty is crucial in order to apply efficient vision-IMU fusion as outlined in IV.

Assume that the positions of the features used to estimate the motion are perfectly known. Then the estimator’s accuracy only depends on the precision of the detected feature locations in the image plane and the spatial configuration of the points with respect to the camera. As bundle adjustment is used on the consecutive frames to compute the precise relative camera pose, the estimated motion parameters already represent a local minimum of the feature reprojection error function.

Let us now consider a point in space, \( \mathbf{x} \), given in homogeneous coordinates in the camera frame. A small transformation \( \mathbf{T}(\Theta) \) is applied to \( \mathbf{x} \) where \( \Theta \) contains the six parameters of the transformation. The transformed point \( \mathbf{T}\mathbf{x} \) can then be projected into the left and right image of the current frame using the projection operations \( L\mathcal{P} \) and \( R\mathcal{P} \). A sample constellation is depicted in Fig. 2.

The reprojection error \( E \) is given by the sum of the squared errors for all points in the left and right image plane

\[
E = \sum_{i=1\ldots N} ||L\mathcal{P}(\mathbf{T}\mathbf{x}_i) - L\mathbf{x}_i||^2 + ||R\mathcal{P}(\mathbf{T}\mathbf{x}_i) - R\mathbf{x}_i||^2
\]  

(1)

\( E \) can now be locally approximated with it’s Hessian \( \mathbf{H} \) as

\[
E(\Theta) \approx E_0 + \frac{1}{2} \Theta^T \mathbf{H} \Theta
\]  

(2)
where $E_0 = E(0)$. Note that the approximation does not contain any linear terms because the reprojection error at the final pose is already a local minimum of $E$ and thus the Jacobian $J$ of $E$ at $\Theta = 0$ is zero.

We can now decompose $E_0$, $J$, and $H$ into the separate terms $E_{0,i}$, $J_i$, and $H_i$ for each of the features. If the feature locations are assumed to be corrupted with additive, white Gaussian noise with variance $\sigma^2$, the probability for a deviation $\Theta$ becomes

$$P(\Theta) \approx \frac{1}{C} \prod_{i=1..N} \exp \left( - \frac{E_{0,i} + J_i \Theta + \frac{1}{2} \Theta^T H_i \Theta}{\sigma^2} \right)$$

$$P(\Theta) \sim \exp \left( - \frac{1}{2\sigma^2} \Theta^T H \Theta \right)$$

$C$ normalizes the distribution. Note that we used $\sum J_i = 0$. From (4) the covariance matrix $\Sigma$ of the pose estimate is identified as

$$\Sigma = \left( \frac{1}{\sigma^2} H \right)^{-1}$$

If degenerate point configurations are encountered, the estimated transformation has remaining degrees of freedom. In this case some Eigenvalues of $H$ will be close to zero and a small regularization $\bar{\Theta}$ is added before inverting $H$.

IV. Extended Kalman Filter

A. Stochastic Cloning

An IMU with accelerometers and rate gyroscopes complements the visual tracking system. It increases the robustness of the visual frame to frame motion estimation and provides pose information if feature tracking fails, e.g. due to occlusions from close objects, bad lighting or ambiguous features. We use an indirect Extended Kalman Filter (EKF) to combine the visual and inertial measurements. However, we cannot incorporate the visual measurements directly in the Kalman framework as they only indicate the relative motion between two time instants instead of absolute measurements. We therefore augment the EKF state with the pose of the previous stereo frame. This method allows us to integrate relative pose estimates between stereo frames and is referred to as stochastic cloning [12], [13].

The state $x$ of our system consists of five quantities: position $p$, rotation $\bar{q}$ as unit quaternion, velocity $v$, gyroscope biases $b$ and accelerometer biases $c$. Further we include $p_d$ and $\bar{q}_d$, the position and rotation of the previous frame.

$$x = \begin{bmatrix} p & q & v & b & c & p_d & \bar{q}_d \end{bmatrix}^T$$

Initially we set $p_d = p$ and $\bar{q}_d = \bar{q}$ and adopt the covariances from the current state variables for the corresponding delayed states.

B. Propagation

In the following the hat and tilde symbols denote estimated and measured quantities respectively. The state is propagated via inertial measurements of acceleration $\hat{\dot{a}}$ and rotation rate $\hat{\dot{\omega}}$. Measurements are corrupted by additive white Gaussian noise with variance $\eta_a$ and $\eta_\omega$. The bias drifts are modelled as random walk processes with noise densities $\eta_b$ and $\eta_c$.

The true states and their estimated counterparts evolve as follows:

$$\dot{p} = v \quad \dot{\bar{q}} = \frac{1}{2} \bar{q} \tilde{\omega}$$

$$\dot{v} = R_{\bar{q}} a + g \quad \dot{\tilde{\omega}} = R_{\bar{q}} (\hat{\dot{\omega}} - \bar{\omega})$$

$$\dot{b} = \eta_b \quad \dot{c} = \eta_c$$

$$\dot{p}_d = 0 \quad \dot{\bar{q}}_d = 0$$

The rotation represented by $\bar{q}$ should be very small and can be approximated as [14]

$$\delta \bar{q} \approx \begin{bmatrix} 1 \end{bmatrix}$$

$\delta \bar{q}$ is the three dimensional rotation error state, keeping the state covariance non-singular while the state representation itself is not minimal due to the quaternion representation.

Let $\bar{F}$ and $G$ be the continuous state transition and noise input matrices respectively. The zeroth order approximation $d \bar{F} = I + \bar{F} \cdot \Delta t$ yields the discrete time state transition matrix with which the state $x_{k-1|k-1}$ and the covariance $P_{k-1|k-1}$ can be propagated to the next time step $k$ using the standard equations

$$\dot{P}_{k|k-1} = d \bar{F} P_{k-1|k-1} d \bar{F}^T + C d Q G^T$$

$$\dot{x}_{k|k-1} = x_{k-1|k-1} + \Delta t \cdot \dot{x}_{k-1|k-1}$$

$\delta Q$ is a $6 \times 6$ zero matrix where $cQ$ contains the noise densities $\eta_a$, $\eta_\omega$, $\eta_b$ and $\eta_c$ on its diagonal.

During all propagations the delayed states and their mutual covariances will remain fixed while the current state variables and covariances evolve via the IMU measurements. Most important the covariances between the current state variables and the delayed states evolve as well during propagation. They essentially represent the knowledge about the relative transformation between these frames and allow us to integrate frame to frame relative measurements from the vision system.

C. Measurement Update

A measurement consists of a six degrees of freedom relative pose estimate between the last and the current stereo frame and its covariance matrix $\Sigma \in \mathbb{R}^{6 \times 6}$. As the position of the last stereo frame is still present in the delayed states $p_d$ and $\bar{q}_d$ we can define the measurement as a function on the state $x$. The translation $p_{\Delta}$ and rotation $\bar{q}_{\Delta}$
between the frames is measured in the last frame’s coordinate system. Expressed as measurement residual we arrive at the measurement function

\[ h(x) = \begin{bmatrix} p_\Delta - R_{\text{arr}}^T (\hat{p} - \bar{p}_d) \\ \bar{q}_\Delta (q_{d}^{-1} \bar{q})^{-1} \end{bmatrix} \]

(12)

After computing the Kalman gain matrix \( K \), the error state estimates are given by \( \hat{X} \). All the states are now updated additively except for the quaternion. According to (8) the latter is updated as \( \bar{q}_{d[k]} = \bar{q}_{d[k-1]} \bar{q}_{\Delta} \). \( \bar{q}_{d[k]} \) needs to be normalized after this step to ensure the orientation quaternion has unit norm. Finally we update the covariance matrix via

\[ P_{k|k} = (I - KH) P_{k|k-1} (I - KH)^T + K \Sigma K^T \]

(13)

After the update step the states \( p \) and \( q \) are copied into \( p_d \) and \( \bar{q}_{d} \) and their covariances are adopted from the original states.

D. Observability of the Global States

The use of only relative measurements inherently makes the position and the yaw rotation states unobservable. Their variances will grow unbounded which will finally prevent the filter from stable long term operation. We chose to insert a pseudo position and yaw measurement of the delayed state whenever it is updated. This enables the filter to maintain approximate global position and yaw state estimates and also leaves the option to integrate absolute global position or rotation measurements without changing our filter design.

V. IMU AIDED FEATURE MATCHING

The onboard inertial measurements allow us to constrain the feature matching between consecutive stereo frames. Propagating the pose using inertial measurements provides a point estimate of the relative pose just before observing a new stereo frame. Through the covariance of the predicted pose, a confidence region for feature observation in the image plane of the next frame is constructed. The search for feature matches can then be restricted to a small region around the expected position of the feature as predicted by the IMU. This increases the robustness of the feature matching significantly, especially if texture is repetitive. Despite the overhead for obtaining the confidence regions for each observed inlier, overall computational complexity is reduced because of a limited search space for the feature matcher and the increased number of inliers in the RANSAC procedure. Typical search spaces are depicted in Fig. 3.

Let \( x \) be the position of a feature in the camera frame after the camera has undergone the motion estimated with the IMU. Its nominal image locations \( ^Lx \) and \( ^Rx \) in both cameras are found via the projection functions \( Lp \) and \( Rp \) as

\[ ^Lx = Lp (T \cdot x) \]
\[ ^Rx = Rp (T \cdot x) \]

(14)

where \( T \) is again a small transformation specified in the camera frame and parameterized by \( \Theta \). For the nominal transformation we have \( T = I_{4 \times 4} \). To obtain the uncertainty of the predicted position we use the Jacobian \( J \) with respect to \( \Theta \). If \( C \) is the covariance of the propagated pose in the local frame the feature covariance matrix \( V \) can be written as

\[ V = JCJ^T \]

(15)

This allows us to restrict the correspondence search to an area where the matching feature is observed with high probability, which increases the inlier ratio for the RANSAC step in complicated scenes. Feature matching with prediction was found to increase the number of correct matches by about 10% over all our datasets. On scenes with repeating textures the number of valid constraints increased by a factor of two to three, which clearly improves the visual estimates. Interestingly, propagation through the IMU couples the uncertainty of the predicted pose with the uncertainty of the gyro and accelerometer biases, implicitly enlarging the search space ellipses during initialization phase or tracking loss.

VI. RESULTS

A. Experiments Setup

The proposed method was evaluated on several trajectories with the stereo head mounted on a quadrotor. Despite
the rotor vibrations the reconstruction performance of the algorithm on flown and hand-held sequences was almost identical. We therefore decided to capture all further datasets, including the ones presented here, hand-held. A mock-up environment of the boiler was set up inside the tracking volume of an external motion capture system to provide accurate ground truth of position and attitude. In addition to a part which was cut out of a real boiler, photographs were stitched together to reproduce the appearance of the boiler walls. For all datasets, stereo frames were captured at a frequency of $10 \text{ Hz}$. The stereo rig is composed of two global shutter machine vision grade CMOS imagers separated by a baseline of $14 \text{ cm}$. Additionally, two high power LEDs were mounted to allow for operation in complete darkness. Signal conditioning and anti-alias filtering of the analog accelerometer and gyroscope signals were performed directly on the IMU. The filtered sensor signals were then captured at $100 \text{ Hz}$. Cameras and IMU are synchronized by a hardware trigger signal. A custom exposure control algorithm was developed to adjust gain and shutter of the imagers as well as onboard lighting intensity.

B. Dataset 1, 'boiler-slow'

A smooth trajectory was followed in front and parallel to the surface of the boiler. Velocities of approximately $0.25 \text{ m/s}$ were reached, at a distance of $70 \text{ cm}$. On this sequence the visual pose estimation uses about 32 features on average to estimate the frame to frame motion. All frames could be tracked visually. The resulting path is plotted in Fig. 4. Table I summarizes the results.

C. Dataset 2, 'boiler-fast'

This dataset tests the estimation during dynamic maneuvers close to the wall. It involves accelerations of up to $6.5 \text{ m/s}^2$ and turn rates reaching $300^\circ/\text{s}$. Only two frames could not be tracked visually. However, $5\%$ of the vision estimates are based on fewer than 10 inliers and on mostly degenerate configurations providing little constraints on the motion. Table I indicates that the accuracy degrades. As expected, the final position error is larger but the orientation and velocity estimates (Fig. 5) are stable and indicate that autonomous operation is feasible even in extreme flight conditions.

D. Dataset 3, 'stairs'

Apart from the boiler mock-up the framework was also tested in general indoor scenarios. Fig. 7 shows the reconstructed path of an indoor trajectory of over $400 \text{ m}$ in length over multiple floors including several staircases. The dataset includes challenging scenes with poor illumination, passing pedestrians, sliding doors and scenes that offer almost no visible texture. The final position error was approximately $0.7\%$ (2.5 m in XY direction and 1 m vertically).

E. Runtime

The target platform for the proposed motion estimation algorithm is a $1.6 \text{ GHz}$ Intel ATOM based single core embedded computer running a standard *Ubuntu 10.04* operating system.
system. Fig. 8 shows the runtime of the different processing steps on the target platform as well as on a standard notebook. Feature extraction is clearly the computational bottleneck. Operations that rely heavily on floating point operations such as bundle adjustment or the Kalman filter are especially slow on the embedded device. The complete odometry system is able to run at approximately 13.5 Hz on the target platform.

VII. CONCLUSION AND FUTURE WORK

Our experiments show that the proposed vision-IMU odometry system is capable of tracking egomotion reliably in the specific environment of the industrial boiler as well as in general indoor scenarios. Reconstruction is accurate despite using only frame to frame feature correspondences. Tightly coupling vision and inertial measurements is effective and improves robustness in situations where purely vision based tracking is difficult. Efficient keypoint detection, descriptor computation and matching enables pose estimation at frame rate on an embedded platform.

To further increase robustness and accuracy, several sub-systems can be optimized. An accurate camera-IMU calibration would certainly increase the accuracy. Currently we only account for camera-IMU translation using measured offsets. Other options include higher quality inertial sensors and multi-frame bundle adjustment, the latter at higher computational cost. We will also consider implementing the most demanding task, keypoint detection, on an FPGA. This will potentially allow us to work with higher resolution imagers and increase the frame rate.

VIII. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2010-2013) under grant agreement n. 248669 (AIRobots). We would also like to thank the Institute for Dynamic Systems and Control [15] at ETH Zürich for providing us access to their Vicor external tracking system.

REFERENCES


