2D-3D Model Correspondence for Camera Pose Estimation using Sensor Fusion

Jeroen Hol and Per Slycke
Xsens Technologies B.V.
P.O. box 545
7500 AM Enschede
The Netherlands
e-mail: {jeroen,per}@xsens.com

Thomas Schön and Fredrik Gustafsson
Division of Automatic Control
Department of Electrical Engineering
Linköping University
SE-581 83 Linköping
Sweden
e-mail: {schon,fredrik}@isy.liu.se

***Abstract*—Augmented reality requires accurate position and orientation estimates of a camera. In our approach inertial and visual data are fused using the conditional probability framework to estimate camera pose. We propose a measurement equation for 2D-3D model correspondence, i.e., detecting known 3D features of the scene in a camera image. The probability density of this measurement equation is investigated and results from a first application are reported.**

***Index Terms*—Augmented reality, 2D-3D model correspondence, Sensor fusion, Vision, Inertial measurements**

I. Introduction

There are many applications in which it is necessary to overlay a computer-generated object onto a real scene in real-time. This requires accurate measurements of the position and orientation (pose) of the camera with respect to the scene.

To compute the pose of the camera, a miniature inertial measurement unit (IMU) attached to the camera is here considered as the main sensor. It consists of three accelerometers and three gyroscopes, which are integrated to obtain pose estimates. However, this leads to a rapid drift in both position and orientation, in particular for MEMS-based miniature inertial sensors, so the IMU needs aiding. A natural choice for the aiding sensor is to use the camera [1]. This option mimic nature: human beings orient themselves using the vestibular organ (in the ears) - which is basically an inertial measurement unit - and the eyes - which are comparable to a camera.

The problem of obtaining real-time pose estimates by fusing both sensors excellently fits the conditional probability framework. Methods from this framework, e.g., Kalman filters [2], [3] and particle filters [4], [5], recursively infer new measurements with knowledge of the system obtained using past measurements. To do so, the dynamic system has to be described, typically in the form of a nonlinear state-space model

$$x_{t+1} = f(x_t, u_t, w_t, t), \tag{1a}$$

$$y_t = h(x_t, e_t, t), \tag{1b}$$

where x_t is the state variable. The state variable contains all necessary information about the system at time t. The process model (1a) describes the evolution of the states x_t given inputs u_t. The observation model (1b) describes the relation between the measurements y_t and the states. Both the process and the observation model can be time-varying, indicated with t in (1). Uncertainty is included by means of the process noise w_t and measurement noise e_t. Realistic models and associated uncertainty are essential for the quality of the estimates.

In this paper, we concentrate on modelling the 2D-3D model correspondence measurement equation (1b) and present preliminary results using an authentic camera motion. These results are compared to an existing reference system.

II. Camera Motion Model

The kinematics of the camera can be described by the following continuous time process model

$$\frac{\partial}{\partial t}\begin{bmatrix}\boldsymbol{c}_f \\ \dot{\boldsymbol{c}}_f \\ q_{sf}\end{bmatrix} = \begin{bmatrix}\dot{\boldsymbol{c}}_f \\ \bar{q}_{sf}\boldsymbol{a}_s q_{sf} + \boldsymbol{g}_f \\ \boldsymbol{\omega}_s q_{sf}\end{bmatrix} \tag{2}$$

where the state vector $\boldsymbol{x} = [\boldsymbol{c}_f^T, \dot{\boldsymbol{c}}_f^T, q_{sf}^T]^T$ consists of camera position $\boldsymbol{c}_f$ and velocity $\dot{\boldsymbol{c}}_f$ and a quaternion [6] q_{sf} describing the orientation of the camera. The acceleration $\boldsymbol{a}_s$ and the angular velocity $\boldsymbol{\omega}_s$ from the IMU are used as input signals. This integration of inertial measurements is known as dead-reckoning. By assuming that the input signals are constant between two sampling instants it is straightforward to derive a discrete time version of (2).

III. Using Vision as a Sensor

The measurements from a vision sensor can be used for camera pose estimation in several ways:

- **2D-3D correspondence**: Positions of fixed point features in the scene and associated uncertainties are contained in a model. These features can be artificial markers placed in the scene, but unobtrusive natural features can also be used. An algorithm that detects a modelled feature in a camera image defines a line where the camera must be with respect to the scene. This relation, here referred to as a 2D-3D model correspondence, can be used to infer the position and orientation of the camera. This type of measurement is analogous to angle only tracking (triangulation). Examples where 2D-3D model correspondence has been previously used are [7], [8].

- **2D-2D correspondence**: Without the scene model a detected point feature or point of interest (POI) by itself is useless. However, if the same POI is detected in sequential images (and the association problem is resolved unambiguously), the displacement of the POI gives information on the (position and angular) velocity [9]. This method can also be applied to lines [10], [11].

In the remainder of this paper 2D-3D model correspondence will be discussed. The image processing required to detect a feature and how to associate 2D-features to 3D-features lies outside the scope of this paper; instead, it focuses on the associated measurement equation and more specifically on its probability distribution.

A. Measurements

The camera image is a projection of the scene on to a plane. For describing this projection the following coordinate systems are used:

- **Fixed (f)**: This is the reference system, fixed to earth. Ignoring the earth's rotation, this is an inertial system. The (static) features of the scene are modelled in this coordinate system.
- **Camera (c)**: The coordinate system attached to the (moving) camera. Its origin is located in the optical centre of the camera.
- **Image (i)**: The 2D coordinate system of camera images. Images are projections of the camera scene into this system.

In this work the projection is modelled by the pinhole camera model [12]. This projection and the coordinate systems involved are illustrated in Fig. 1. Although the pinhole camera model is somewhat simplistic to be realistic, the resulting equations can be adapted to more complex camera models. The projection of a feature from the camera coordinate system to the image coordinate system, $\boldsymbol{z}_c \mapsto \boldsymbol{\zeta}_i$, with $\boldsymbol{z}_c = [x, y, z]^T$ and $\boldsymbol{\zeta}_i = [\xi, \psi]^T$, is given by

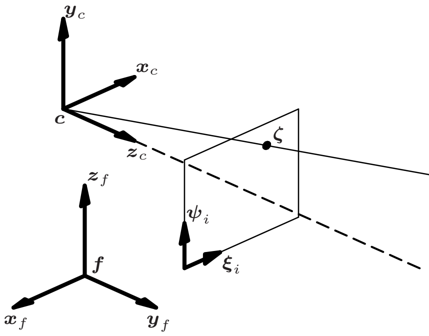


Fig. 1. Illustration of the different coordinate systems and how they are related. Furthermore, the projection $\boldsymbol{\zeta}$ of feature $\boldsymbol{z}$ is shown in detail.

$$\left[\xi, \psi\right]^T = \left[fx/z, fy/z\right]^T \tag{3}$$

Expressed in projective space, as is customary within the field of image processing, this relation is

$$\left[\xi, \psi, 1\right]^T \cong \left[fx, fy, z\right]^T \tag{4}$$

where $\cong$ means equality up to scale. This scale can be eliminated from the equation by applying the cross product operator, that is

$$\boldsymbol{a} = \lambda \boldsymbol{b} \Leftrightarrow \boldsymbol{a} \times \boldsymbol{b} = \lambda \boldsymbol{b} \times \boldsymbol{b} = 0 \tag{5}$$

resulting in

$$\begin{bmatrix} \xi \\ \psi \\ 1 \end{bmatrix} \times \begin{bmatrix} fx \\ fy \\ z \end{bmatrix} = \begin{bmatrix} z\psi - fy \\ z\xi - fx \\ f(\xi y - \psi x) \end{bmatrix} = 0 \tag{6}$$

Note that the first two components of this vector are very similar to (3). The last component is redundant, as only two image coordinates are measured. From a probabilistic point of view (3) is not preferred as the division of two normal variable results in a Cauchy distribution, which has infinite second and higher order moments. Hence, (6) will be used in the subsequent sections.

B. Probability Density Function

Defining a virtual measurement $\boldsymbol{h}$ according to (6)

$$\boldsymbol{h}(\boldsymbol{z}_c, \boldsymbol{\zeta}_i) = \begin{bmatrix} z\xi - fx \\ z\psi - fy \end{bmatrix} \tag{7}$$

its probability density function (PDF), $p_h(h_1, h_2)$, is given by

$$\iint \delta(z\xi - fx - h_1)\, \delta(z\psi - fy - h_2) p_z(\boldsymbol{z}) p_\zeta(\boldsymbol{\zeta}) d\boldsymbol{z} d\boldsymbol{\zeta} \tag{8}$$

where $\delta(\cdot)$ is Dirac's delta function. Independent and normal distributions are assumed for $\xi \sim \mathrm{N}(\mu_\xi, \sigma_\xi^2)$, $\psi \sim \mathrm{N}(\mu_\psi, \sigma_\psi^2)$ and $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{\mu}_z, \Sigma_z)$. Unfortunately, no analytical expression can be found for (8). Numerical integration is possible, but not feasible due to the very high number of evaluations required by an application. The characteristic function [13]

$$\phi_h(u_1, u_2) = \mathrm{E}[\mathrm{e}^{i\boldsymbol{u}\cdot\boldsymbol{h}}] = \int \mathrm{e}^{i\boldsymbol{u}\cdot\boldsymbol{h}}\, p_h(\boldsymbol{h}) d\boldsymbol{h} \tag{9}$$

can be derived. From (9) all moments can be calculated by

$$\mathrm{E}[h_1^n h_2^m] = i^{-(n+m)}\, \frac{\partial^{n+m}}{\partial u_1^n \partial u_2^m} \phi_h(\boldsymbol{u}) \bigg|_{\boldsymbol{u}=0} \tag{10}$$

The first two central moments of $\boldsymbol{h}$, its mean and covariance, are given by

$$\mathrm{E}[\boldsymbol{h}] = \begin{bmatrix} \mu_z \mu_\xi - f\mu_x \\ \mu_z \mu_\psi - f\mu_y \end{bmatrix} \tag{11a}$$

$$\begin{aligned} \mathrm{E}[\bar{h}_1^2] &= f^2 s_{xx} - 2f\mu_\xi s_{xz} + \mu_\xi^2 s_{zz} + (\mu_z^2 + s_{zz}) s_{\xi\xi} \\ \mathrm{E}[\bar{h}_1 \bar{h}_2] &= f^2 s_{xy} - f\mu_\psi s_{xz} - f\mu_\xi s_{yz} + \mu_\xi \mu_\psi s_{zz} \\ \mathrm{E}[\bar{h}_2^2] &= f^2 s_{yy} - 2f\mu_\psi s_{yz} + \mu_\psi^2 s_{zz} + (\mu_z^2 + s_{zz}) s_{\psi\psi} \end{aligned} \tag{11b}$$

where $\bar{\boldsymbol{h}} = \boldsymbol{h} - \boldsymbol{\mu}_h$, $\mu_a = \mathrm{E}[a]$ and $s_{ab} = \mathrm{Cov}(a, b)$. Note the similarity between (11a) and (7). The third order central moments of $\boldsymbol{h}$ are given by

$$\begin{aligned} \mathrm{E}[\bar{h}_1^3] &= 6\mu_z \mu_\xi s_{zz} s_{\xi\xi} - 6\mu_z f s_{xz} s_{\xi\xi} \\ \mathrm{E}[\bar{h}_1^2 \bar{h}_2] &= 3\mu_z \mu_\psi s_{zz} s_{\xi\xi} - 3\mu_z f s_{yz} s_{\xi\xi} \\ \mathrm{E}[\bar{h}_1 \bar{h}_2^2] &= 3\mu_z \mu_\xi s_{zz} s_{\psi\psi} - 3\mu_z f s_{xz} s_{\psi\psi} \\ \mathrm{E}[\bar{h}_2^3] &= 6\mu_z \mu_\psi s_{zz} s_{\psi\psi} - 6\mu_z f s_{yz} s_{\psi\psi} \end{aligned} \tag{12}$$

The variances of the image and model can be assumed small in a practical situation, hence products of covariance terms are small, i.e. $s^2 \ll 1$. This implies that the moments in (12) are negligible small, so a normal approximation with mean and covariance according to (11) is correct up to 3rd order. A comparison with a numerical solution, see Fig. 2 for a typical example, shows almost identical densities. Therefore, it is argued that, for practical applications, a normal approximation can be used to adequately describe the PDF of $\boldsymbol{h}$.

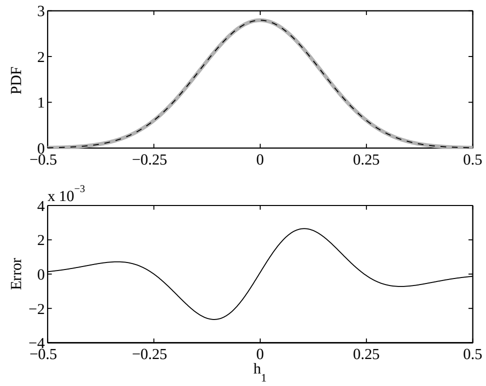


Fig. 2. The marginal PDF of h_1 (solid) compared to its normal approximation (dashed).

C. Measurement Equation

The features positions in the scene are expressed using the fixed coordinate system. Hence, these positions have to be transformed into the camera coordinate system in order to use (7). This transformation $\boldsymbol{z}_f \mapsto \boldsymbol{z}_c$ has the form

$$\boldsymbol{z}_c = R_{cf}[\boldsymbol{z}_f - \boldsymbol{c}_f] \tag{13}$$

where $\boldsymbol{c}_f$ is the position of the camera in the fixed coordinate system and R_{cf} is the rotation matrix from the fixed system to the camera system. The coordinate system transformation affects the covariance matrix as well,

$$\Sigma_{z_c} = R_{cf}\Sigma_{z_f}R_{cf}^T \tag{14}$$

Substituting these transformations into (7) results in an implicit measurement equation for 2D-3D model correspondence

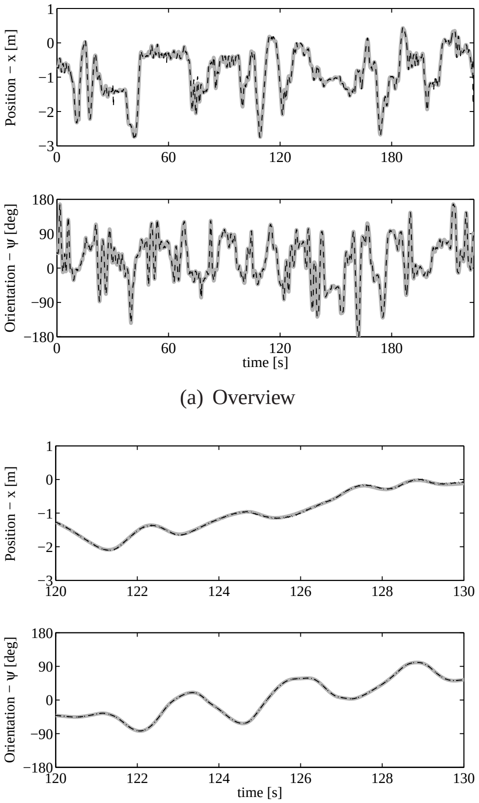
$$0 = \boldsymbol{h}(R_{cf}[\boldsymbol{z}_f - \boldsymbol{c}_f], \boldsymbol{\zeta}_i) + \boldsymbol{e} \tag{15}$$

with $\boldsymbol{e} \sim \mathrm{N}(0, \Sigma_h)$ and Σ_h given by (11b).

IV. Application

The 2D-3D model correspondence measurement equation has been applied to track a camera held by a running cameraman, which is moving rapidly, hence it is quite hard to track. The process of tracking here refer to estimating the position and orientation. The camera is fitted with a Xsens MT9 IMU which provides acceleration and angular velocity at 100 Hz. A camera pose reference is provided by the free-D system [14], which allows to benchmark tracking accuracy. For testing purposes an artificial scene model, consisting of a sphere of 100 features of which on average 14 are visible, is used. The image measurements were generated at 25 Hz by projecting the scene model using the reference camera pose. Both the 2D and 3D feature data were corrupted using white noise with standard deviations of 10^{-3} m and 10^{-2} m respectively.

The camera pose has been estimated using an Extended Kalman Filter (EKF) [3]. This filter use the measured inertial data as input signals in the process model and the artificial visual data as measurements. Fig. 3(a) shows typical results for position and orientation. Only one horizontal position (x) and orientation (heading ψ) are shown, the other positions and orientations show similar behaviour. Fig. 3(a) clearly shows that tracking of rapid and large movements is possible using 2D-3D model correspondence. Looking in more detail to the results, see Fig. 3(b), one sees that the orientation is very accurate. The position responds a little late. This delay is likely to be the result of the poorly observable velocities inherent to using only 2D-3D model correspondence measurements.



(a) Overview

(b) Detail

Fig. 3. Pose estimates using 2D-3D model correspondence and inertial measurements (dashed) compared to reference system (solid) for a running cameraman. From the 6 DOFs only horizontal position (x) and heading (ψ) are shown.

The next step will be to use image processing for the visual measurements and improve the inertial sensor error model.

V. Conclusion

In this paper a measurement equation for 2D-3D model correspondence is proposed together with an evaluation based on realistic data. The probability density function of the measurement equation could not be calculated analytically, but it is shown that that it can be adequately approximated by a normal probability density function. This approximation has correct central moments up to 3rd order and a comparison with numerical approximations shows almost identical probability densities. These results provide a theoretical foundation for using 2D-3D model correspondence for pose estimation within the conditional probability framework and, more specifically, for usage in (Extended) Kalman filters.

The derived measurement equation has been implemented in an Extended Kalman Filter using authentic inertial measurements and simulated 2D-3D model correspondence measurements. Comparing the results to the reference system shows stable and accurate position and orientation tracking over an extended period of time for camera that undergoes fast motion.

Acknowledgement

This work has been performed within the MATRIS consortium, which is a sixth framework research program within the European Union (EU), contract number: IST-002013. The authors would like to thank the EU for the financial support and the partners within the consortium for a fruitful collaboration this far. The MATRIS consortium aims to develop a marker-free tracking system. For more information, please visit its website, `www.ist-matris.org`.

References

[1] R. Azuma, “A survey of augmented reality,” *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 4, pp. 355–385, August 1997.

[2] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Trans. AMSE, J. Basic Engineering*, vol. 82, pp. 35–45, 1960.

[3] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*, ser. Information and System Sciences Series. Upper Saddle River, New Jersey: Prentice Hall, 2000.

[4] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*. Springer Verlag, 2001.

[5] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for Bayesian filtering,” *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[6] M. Shuster, “A survey of attitude representations,” *The Journal of the Astronautical Sciences*, vol. 41, no. 4, pp. 439–517, October 1993.

[7] A. Davison, “Real-time simultanious localisation and mapping with a single camera,” in *Proc. International Conference on Computer Vision*, Nice, France, Oct. 2003, pp. 1403–1410.

[8] Y. Yokokohji, Y. Sugawara, and T. Yoshikawa, “Accurate image overlay on video see-through hmds using vision and accelerometers,” in *Proc. IEEE Virtual Reality 2000 Conference*, New Brunswick, USA, March 2000.

[9] T. Schön and F. Gustafsson, “Integrated navigation of cameras for augmented reality,” in *Proceedings of the 16th IFAC world Congress*, Prague, Czech Republic, Jul. 2005, to appear.

[10] H. Rehbinder and B. Ghosh, “Pose estimation using line-based dynamics vision and inertial sensors,” *IEEE Transactions on Automatic Control*, vol. 48, no. 2, pp. 186–199, February 2003.

[11] Y. Wu, M. Wu, D. Hu, and T. Wu, “Observability analysis of rotation estimation by fusing inertial and line-based visual information: A revisit,” *Submitted to IEEE Trans. on Aerospace and Electronic Systems*, 2004.

[12] R. Hartley and A. Zisserman, *Multiple view Geometry in Computer Vision*. Cambridge University Press, 2000.

[13] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 2nd ed. McGraw-Hill, 1984.

[14] G. Thomas, J. Jin, N. Niblett, and C. Urquhart, “A versatile camera position measurement system for virtual reality tv production,” in *International Broadcasting Conference*, Amsterdam, The Netherlands, September 1997, pp. 284–289.