On-line Structure and Motion Estimation based on an Novel Parameterized Extended Kalman Filter

Sebastian Haner and Anders Heyden
Centre for Mathematical Sciences
Lund University, Sweden
Email: {haner, heyden}@maths.lth.se

***Abstract*—Estimation of structure and motion in computer vision systems can be performed using a dynamic systems approach, where states and parameters in a perspective system are estimated. We present a novel on-line method for structure and motion estimation in densely sampled image sequences. The proposed method is based on an extended Kalman filter and a novel parameterization. We assume calibrated cameras and derive a dynamic system describing the motion of the camera and the image formation. By a change of coordinates, we represent this system by normalized image coordinates and the inverse depths. Then we apply an extended Kalman filter for estimation of both structure and motion. The performance of the proposed method is demonstrated in both simulated and real experiments. We furthermore compare our method to the unified inverse depth parameterization and show that we achieve superior results.**

I. Introduction

Estimation of 3D-structure and motion from 2D images is a central problem in computer vision. There exist essentially two different approaches to solve this problem; (i) batch approaches and (ii) iterative (recursive) approaches. Batch approaches aim at providing an accurate result by using all the images at the same time. These approaches are typically based on multi-view tensors, bundle adjustment or convex optimization, see [10] for the former and [12] for the latter. Iterative (or recursive) approaches aim at real-time performance, by updating a current estimate as soon as a new image becomes available. These approaches are either based on variations of methods used for batch approaches, e.g. iteratively estimating the camera pose and the structure, [3], or by fast estimation of relative motion [17].

Yet another approach is to formulate the camera motion and the imaging process as a dynamic system and apply non-linear observers to estimate the structure and the translational and rotational velocities of the motion. The standard approach is to apply an extended Kalman filter to a dynamic system, with a perspective transformation in the output equations. One of the pioneering approaches is [2] where an extended Kalman filter is applied directly to the dynamic system, without any re-parameterization. Another approach, based on tracking the essential matrix can be found in [18].

For structure estimation only, i.e. known motion, a number of non-linear observers based on methods for automatic control theory have been developed, e.g. [1], [5], [7], [9], [13], [14], [16]. Similar approaches, based on adaptive non-linear observers, for full structure and motion estimation can be found in [11], [19], [20].

Lately, [6], [8] developed a variant of the extended Kalman filter, by using the inverse depth as one of the parameters, adjusting the uncertainties to the imaging situation and fixing the imaging rays from the first camera in order to gain stability. The method is highly over-parameterized but performs well in most situations, both in terms of accuracy and robustness. Another approach based on inverse scaling can be found in [15].

This paper describes how a re-parameterization of the underlying perspective dynamic system can be used to formulate the structure and motion estimation problem as an observer problem of a non-linear dynamic system, with a linear output function. We will show that this novel parameterization results in a more accurate and stable extended Kalman filter.

II. Problem formulation

Consider a calibrated perspective camera that is observing a moving rigid object. The camera system can be written as (assuming the camera is situated at the origin and that the optical axis is aligned with the z-axis)

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{x_1}{x_3} \\ \frac{x_2}{x_3} \end{bmatrix} \ , \tag{1}$$

where y_i denote the image coordinates (compensated for intrinsic parameters) and x_i denote the (time-varying) object coordinates. Introducing

$$\xi = \begin{pmatrix} \frac{x_1}{x_3} & \frac{x_2}{x_3} \end{pmatrix}^T \ , \tag{2}$$

we can write down a dynamic system

$$\begin{aligned} \dot{x} &= Ax + b \, , \\ y &= \xi^i \ , \end{aligned} \tag{3}$$

where

$$A = S(\omega) = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix} \tag{4}$$

is the skew symmetric matrix obtained from the (possibly time varying) angular velocity vector

$$\omega = \begin{pmatrix} \omega_1 & \omega_2 & \omega_3 \end{pmatrix}^T \tag{5}$$

and

$$b = \begin{pmatrix} b_1 & b_2 & b_3 \end{pmatrix}^T \ . \tag{6}$$

denote the (possibly time varying) translational velocity. We can now state the problem as follows:

Problem 1 (On-line structure and motion estimation)**.** *Given the image coordinates y from (3), estimate recursively the object coordinates x and the motion parameters ω and b.*

III. The parameterization

Consider (3) and introduce the scalar parameter γ and the vector z,

$$\gamma = \frac{1}{\sqrt{x^T x}}, \quad z = \gamma x \ , \tag{7}$$

where γ can be interpreted as the inverse distance to the object. Observe that ξ, according to (2) and by the definition of z in (7), also can be expressed as

$$\xi = \begin{pmatrix} \frac{z_1}{z_3} & \frac{z_2}{z_3} \end{pmatrix}^T \ . \tag{8}$$

Using (7) and the definition of ξ in (2), the vector z may be expressed as

$$z = \frac{1}{\sqrt{\xi_1^2 + \xi_2^2 + 1}} \begin{pmatrix} \xi_1 & \xi_2 & 1 \end{pmatrix}^T \tag{9}$$

and can thus be assumed known. This vector can be interpreted as the image coordinates on a spherical image plane.

Hence, z is a measurable signal, and can therefore be considered an output of the system (3). The parametrization exploits this fact, and aims at rewriting the system (3) so that z appears explicitly in the equations.

Using (3) and the fact that $x^T A x = 0$ since A is skew-symmetric, gives, introducing

$$g_0(z) = I - zz^T \tag{10}$$

a rewritten dynamic system, corresponding to (3), on the form

$$\begin{aligned} \dot{z} &= Az + g_0(z) b\gamma \\ \dot{\gamma} &= -\gamma^2 z^T b \ . \end{aligned} \tag{11}$$

For the motion of more than one point a dynamic system corresponding to (11) is obtained as

$$\begin{aligned} \dot{z}^i &= Az^i + g_0(z^i) b\gamma^i \\ \dot{\gamma}^i &= -(\gamma^i)^2 (z^i)^T b \end{aligned} \quad , \quad i \in \{1, 2 \dots N\} \ , \tag{12}$$

where N denotes the number of feature points. Equation (7) together with (11) and its multipoint version (12), constitute the desired dynamic vision parameterization, from which we shall proceed. Observe the the dynamic system contains 4 state variables; 3 for z and 1 for γ and that z has to fulfill the constraint $|z| = 1$.

IV. The extended Kalman filter

The extended Kalman filter estimates the system state s_k given a previous estimate $\hat{s}_{k-1}$, a new measurement μ and state transition and observation models $s_k = f(s_{k-1})$ and $\mu_k = h(s_k)$. At every timestep the new state and the state covariance P are predicted,

$$\begin{aligned} \hat{s}_{k|k-1} &= f(\hat{s}_{k-1|k-1}) \\ P_{k|k-1} &= F_{k-1} P_{k-1|k-1} F_{k-1}^T + Q_{k-1} \end{aligned} \tag{13}$$

and, given a new measurement μ_k, corrected to

$$\begin{aligned} \hat{s}_{k|k} &= \hat{s}_{k|k-1} + K_k\big(\mu_k - h(\hat{s}_{k|k-1})\big) \\ P_{k|k} &= P_{k|k-1} - K_k H_k P_{k|k-1} \end{aligned} \tag{14}$$

where

$$\begin{aligned} F_{k-1} &= \left.\frac{\partial f}{\partial s}\right|_{\hat{s}_{k-1|k-1}}, \quad H_k = \left.\frac{\partial h}{\partial s}\right|_{\hat{s}_{k|k-1}} \\ K_k &= P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + R_k)^{-1} \end{aligned} \tag{15}$$

and Q and R the assumed process and measurement noise covariances, respectively.

Adapting the dynamic system (12) to the EKF setting, the state vector is taken to be

$$s = [b^T, \omega^T, (z^1)^T, \gamma^1, \dots, (z^N)^T, \gamma^N]^T, \tag{16}$$

and measurements

$$\mu = [\xi_1^1, \xi_2^1, \dots, \xi_1^N, \xi_2^N]^T. \tag{17}$$

The measurement equation is simply

$$[\xi_1^i, \ \xi_2^i] = \Big[\frac{z_1^i}{z_3^i}, \ \frac{z_2^i}{z_3^i}\Big], \tag{18}$$

while the update equation is a discretized version of (12):

$$\begin{aligned} \tilde{z}^i &= e^{S(\omega_k)} z_k^i + g_0(z_k^i) b_k \gamma_k^i \\ \tilde{\gamma}^i &= \gamma_k^i - (\gamma_k^i)^2 (z_k^i)^T b_k \\ z_{k+1}^i &= \tilde{z}^i |\tilde{z}^i|^{-1} \\ \gamma_{k+1}^i &= \tilde{\gamma}^i |\tilde{z}^i| \end{aligned} \quad , \quad i \in \{1, 2 \dots N\} \ , \tag{19}$$

where also $|z^i| = 1$ is enforced. An alternative approach is to store only two of the components of z in the state vector, and reconstruct the third, when needed, as $z_3 = \sqrt{1 - z_1^2 - z_2^2}$, obviating the need for normalization above. While the EKF is not guaranteed to keep $z_1^2 + z_2^2 \leq 1$, in practice it is found that this works well, and has the benefit of minimizing the state vector size.

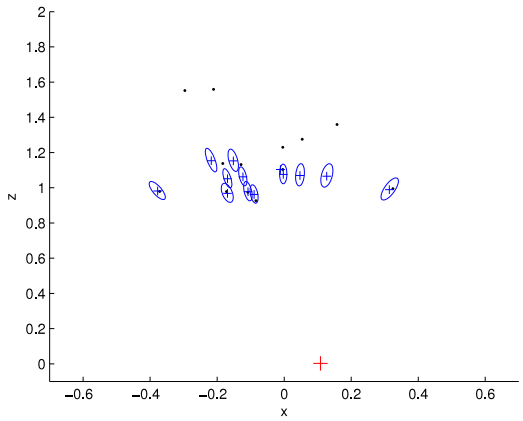
Note that we assume a camera-centric coordinate system and estimate only linear and angular velocities, which must be integrated over time to recover the absolute motion. The approach is similar to the one presented in [4] which is a camera-centric version of the unified inverse depth method [6].

V. Experiments

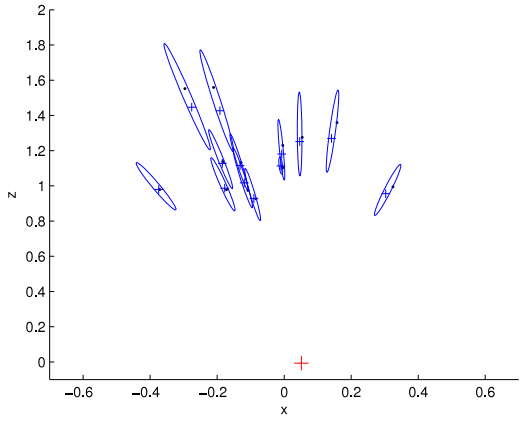
In the following experiments, no priors on the structure or motion are given. Features are initialized at an arbitrary depth and with large uncertainty in the γ coordinate. The linear and angular velocities are assumed constant, and acceleration is modeled as zero-mean Gaussian process noise.

As has been reported in [6], the EKF can converge under these circumstances; however, it is found that fixing the depth of one point, thus determining the overall scale, greatly aids convergence. Further, the normalization step of the update equation (19) has been found not strictly necessary (when using the full parameterization) and in fact does not significantly impact the results.

We repeat an experiment in [15] and show that the proposed parameterization does not suffer from the underestimation of uncertainty associated with the inverse depth parameterization of [6] and typically converges faster as a result (figures 1 and 2).



(a) Unified inverse depth



(b) Proposed

Fig. 1. Position and covariance estimates after observing 30 frames of simulated data (black: ground truth, blue: estimate $\pm\sigma$). The inverse depth parameterization underestimates the errors, here leading to slower convergence, while the proposed parameterization more accurately captures the depth uncertainty.

The proposed parameterization shares the measurement linearity with the inverse depth parameterization, since they are equivalent in the limit of small angles. Due to the camera-centric representation, however, the update equation is not linear.

We also apply the proposed and unified inverse depth methods to a real video sequence, reconstructing camera motion and 3D coordinates of 7 feature points tracked over 70 frames. Some geometry is overlaid to verify the results (figure 3). A (subjective) assessment indicates that the proposed method gives a more consistent reconstruction.

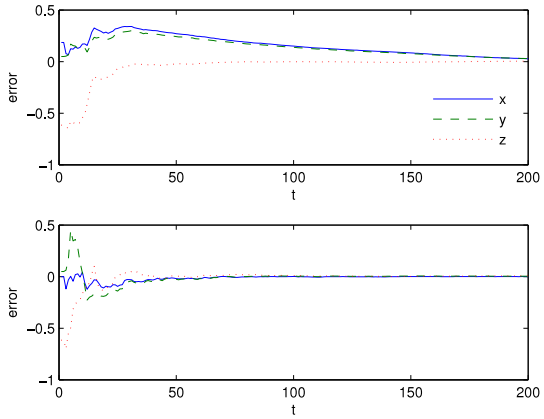


Fig. 2. Convergence plot of the Cartesian coordinates of a point in a simulated reconstruction problem. Top: inverse depth, bottom: proposed parameterization.

VI. Conclusions

We have used a novel parameterization in order to develop an extended Kalman filter for full structure and motion estimation. The filter is shown to perform well on both simulated and real data and has been compared to the current state-of-the-art. Further studies will be to handle missing data, e.g. novel and disappearing tracks, and increase the robustness to false matches.

Acknowledgements

The authors would like to acknowledge financial support form the SRC-grant 621-2008-4557.

References

[1] R. Abdursul, H. Inaba, and B. K. Ghosh, "Nonlinear observers for perspective time-invariant linear systems," *Automatica*, vol. 40, pp. 481–490, 2004.

[2] A. Azarbayejani and A. P. Pentland, "Recursive estimation of motion, structure, and focal length," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 6, pp. 562–575, 1995.

[3] P. Beardsley, A. Zisserman, and D. Murray, "Sequential updating of projective and affine structure from motion," *International Journal of Computer Vision*, vol. 23, no. 3, pp. 235–259, 1997.

[4] A. Boberg, A. N. Bishop, and P. Jensfelt, "Robocentric mapping and localization in modified spherical coordinates with bearing measurements," *ISSNIP*, 2009.

[5] X. Chen and H. Kano, "A new state observer for perspective systems," *IEEE Transactions on Automatic Control*, vol. 47, no. 4, pp. 658–663, April 2002.

[6] J. Civera, A. Davison, and J. Montiel, "Inverse depth parametrization for monocular slam," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 932–945, 2008.

[7] O. Dahl, F. Nyberg, and A. Heyden, "Nonlinear and adaptive observers for perspective dynamic systems," in *American Control Conference*, July 2007.

[8] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[9] W. E. Dixon, Y. Fang, D. M. Dawson, and T. J. Flynn, "Range identification for perspective vision systems," *IEEE Transactions on Automatic Control*, vol. 48, no. 12, pp. 2232–2238, December 2003.

[10] R. Hartley and A. Zisserman, *Multiple View Geometry*. Cambridge University Press, 2003.

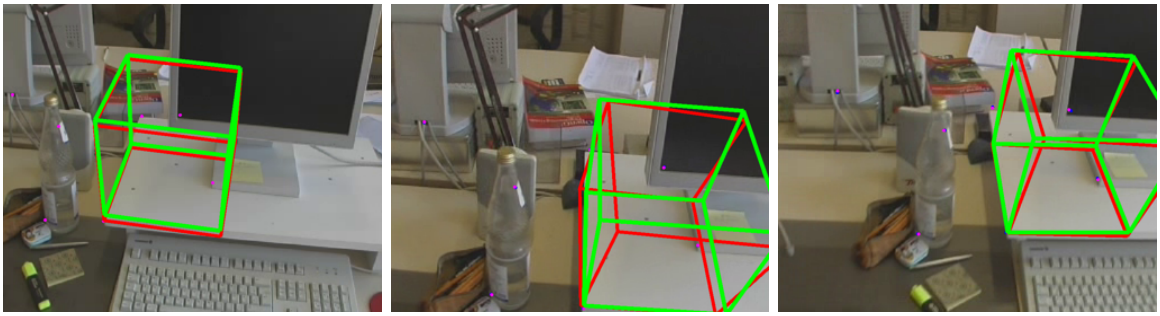


Fig. 3. Visual result of integrating geometry into a tracked video sequence (from left to right, frames 1, 50 and 70 are shown). The green box shows the solution using the proposed method, while the red was computed using the inverse depth parameterization. Although the reprojection errors are similar, the proposed method produces a more accurate motion estimate.

[11] A. Heyden and O. Dahl, "Provably convergent structure and motion estimation for perspective systems," in *Control Decision Conference*, 2009.

[12] F. Kahl, "Multiple view geometry and the L_∞-norm," in *International Conference on Computer Vision*. IEEE Computer Society Press, 2005, pp. 1002–1009.

[13] D. Karagiannis and A. Astolfi, "A new solution to the problem of range identification in perspective vision systems," *IEEE Transactions on Automatic Control*, vol. 50, no. 12, pp. 2074–2077, December 2005.

[14] L. Ma, Y. Chen, and K. L. Moore, "Range identification for perspective dynamic systems with 3d imaging surfaces," in *American Control Conference*, June 2005.

[15] D. Marzorati, M. Matteucci, D. Migliore, and D. Sorrenti, "Monocular slam with inverse scaling parametrization," in *BMVC08*, 2008.

[16] L. Matthies, T. Kanade, and R. Szeliski, "Kalman filter-based algorithms for estimating depth from image sequences," *International Journal of Computer Vision*, vol. 3, pp. 209–236, 1989.

[17] D. Nister, "An efficient solution to the five-point relative pose problem," *Int. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 195–202, 2003.

[18] S. Soatto, "3-d structure from visual motion: Modeling, representation and observability," *Automatica*, vol. 33, no. 7, pp. 1287–1312, 1997.

[19] S. Soatto and P. Perona, "Reducing structure from motion: A general framework for dynamic vision, part 1: Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, 1998.

[20] ——, "Reducing structure from motion: A general framework for dynamic vision, part 2: Implementation and experimental assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, 1998.