Dimensionless Monocular SLAM

Javier Civera[1], Andrew J. Davison[2], and J. M. M. Montiel[1]

[1] Dpto. de Informática e Ingeniería de Sistemas, University of Zaragoza, Spain
{jcivera, josemari}@unizar.es
[2] Department of Computing, Imperial College, London, UK.
ajd@doc.ic.ac.uk

Abstract. It has recently been demonstrated that the fundamental computer vision problem of structure from motion with a single camera can be tackled using the sequential, probabilistic methodology of monocular SLAM (Simultaneous Localisation and Mapping). A key part of this approach is to use the priors available on camera motion and scene structure to aid robust real-time tracking and ultimately enable metric motion and scene reconstruction. In particular, a scene object of known size is normally used to initialise tracking.
In this paper we show that real-time monocular SLAM can be initialised with no prior knowledge of scene objects within the context of a powerful new dimensionless understanding and parameterisation of the problem. When a single camera moves through a scene with no extra sensing, the scale of the whole motion and map is not observable, but we show that up-to-scale quantities can be robustly estimated.
Further we describe how the monocular SLAM state vector can be partitioned into two parts: a dimensionless part, representing up-to-scale scene and camera motion geometry, and an extra metric parameter representing scale. The dimensionless parameterisation permits tuning of the probabilistic SLAM filter in terms of image values, without any assumptions about scene scale, but scale information can be put back into the estimation if it becomes available.
Experimental results with real image sequences showing SLAM without an initialisation object, different image tuning examples and scenes with the same underlying dimensionless geometry are presented.

1 Introduction

Structure From Motion (SFM) [5], classically solved as batch process, has recently been reformulated as a sequential probabilistic estimation problem, propagating and benefitting from available priors along an image sequence. The probabilistic approach is based on SLAM techniques from the mobile robotics field, using either the Extended Kalman Filter (EKF) [3, 6] or particle filtering methods such as FastSLAM [4]. This rigorous Bayesian approach is producing a significant improvement both in matching robustness and computation speed. Systems built using commodity cameras and computers have shown real-time 30 fps. robust performance in indoor or outdoors scenes with a hand-held camera.

It is a well known fact in SFM that a moving calibrated camera observing a scene can recover scene geometry and camera motion only up to a scale factor — scene scale is an non-observable magnitude if only bearing measurements are made. Unlike SFM, probabilistic SLAM methods use prior information: a camera motion model, scene depth priors and some *known structure*. These priors both aid sequential tracking (by defining search regions) and enable the computation of a *metric scene scale*. In particular, current monocular SLAM methods [3, 6] have used extra information in the form of a known initialisation object to fix scene scale.

In this paper we show that this non-visual information is in fact not essential for solving the tracking problem and that no known target object needs to be added to the scene. While this means that overall scene scale cannot intrinsically be recovered, real-time tracking can still proceed — and if extra information does become available later, scale can be put back into the scene map.

This is enabled by a novel understanding of the monocular SLAM problem, based on the Extended Kalman Filter (EKF), in terms of dimensionless parameters. The new parameterisation is derived using Buckingham's Π theorem [1] which relies on the necessity for dimensional correctness in any formula and hence any estimation process. Our monocular SLAM algorithm therefore recovers dimensionless, up-to-scale geometry, and also provides benefits by allowing previous tuning parameters to be rolled up into a canonical set which give an important new understanding of the uncertainties in the system now in pixel units. These parameters in the image provide a natural way of understanding image sequences, irrespectively of the frame rate and actual scene size.

Further, we show that alongside the main dimensionless part of the SLAM state vector we can add an extra parameter representing metric scale. During tracking, vision-only measurements do not reduce the uncertainty in the scale parameter but only in the dimensionless scene geometry. However, any measurement containing metric information such as odometry, a feature at a known depth or the distance between two features can be added when available and will correctly affect both the scale and the dimensionless scene geometry.

2 Monocular SLAM Estimation Process

The state of the system in EKF SLAM is traditionally represented by a state vector $\mathbf{x}$, composed of a group of parameters referring to the camera motion, $\mathbf{x_v}$, and n others representing every feature in the map, $\mathbf{y_i}$ [7, 2].

$$\mathbf{x} = (\mathbf{x_v}, \mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n})^\top \tag{1}$$

In hand-held camera monocular SLAM, a smooth camera motion is usually supposed. The motion model in this paper is the same as in [3]: a constant velocity model with unknown acceleration inputs, a_k^W and α_k^C. These linear and angular accelerations are represented by zero mean known standard deviations (σ_a and σ_α) Gaussian noise . The camera state vector includes camera location,

rotation quaternion, and linear and angular velocities:

$$\mathbf{x_v} = \left(\mathbf{r}^{WC}, \mathbf{q}^{WC}, \mathbf{v}^{W}, \omega^{W}\right) \tag{2}$$

The equation that updates the state camera vector at every step is:

$$\mathbf{f_v} = \begin{pmatrix} \mathbf{r}^{WC}_{k+1} \\ \mathbf{q}^{WC}_{k+1} \\ \mathbf{v}^{W}_{k+1} \\ \omega^{C}_{k+1} \end{pmatrix} = \begin{pmatrix} \mathbf{r}^{WC}_{k} + v^{W}_{k}\Delta t + a^{W}_{k}\Delta t^2 \\ \mathbf{q}^{WC}_{k} \times \mathbf{q}(\omega^{C}_{k}\Delta t + \alpha^{C}_{k}\Delta t^2) \\ \mathbf{v}^{W}_{k} + a^{W}_{k}\Delta t \\ \omega^{C}_{k} + \alpha^{C}_{k}\Delta t \end{pmatrix} \tag{3}$$

Inverse depth parametrization for point features [6] is also used in this paper. This parametrization codes features by the ray extracted at first feature observation (defined by the 3D location of the optical centre of the camera and azimuth-elevation angles) and the inverse depth along this ray:

$$\mathbf{y_i} = (\mathbf{r}_i, \theta_i, \phi_i, \rho_i) = (x_i, y_i, z_i, \theta_i, \phi_i, \rho_i) \tag{4}$$

When a feature is newly initialized from a monocular camera, only information about the ray can be retrieved. As no information is available about depth, an initial inverse depth Gaussian prior on $\rho_i \sim N(\rho_0, \sigma_{\rho_0})$ is applied in order to cover with 95% probability the range of depths from the closest possible to infinity.

We propose to split the state vector into a metric parameter d — unobservable with only-vision measurements — and a dimensionless scene and camera part. Doing this, the state vector is partitioned according to observability with a monocular camera. Camera measurements will reduce scene geometry uncertainty, but not the uncertainty in the metric parameter d.

$$\mathbf{x} = \left(d, \mathbf{\Pi^{WC}_r}, \mathbf{q^{WC}}, \mathbf{\Pi^{W}_v}, \mathbf{\Pi^{C}_\omega}, \mathbf{\Pi_{y_1}}, \ldots\right)^\top \tag{5}$$

The mapping from the state vector to metric scene geometry is a non-linear computation involving the dimensionless geometry and the parameter d:

$$\mathbf{r}^{WC} = d\mathbf{\Pi^{WC}_r}, \qquad \mathbf{v}^{W} = d\mathbf{\Pi^{W}_v}\Delta t, \qquad \omega^{W} = d\mathbf{\Pi^{W}_\omega}\Delta t \tag{6}$$

$$\mathbf{y_i} = (d\Pi_{xi}, d\Pi_{yi}, d\Pi_{zi}, \theta_i, \phi_i, \Pi_{\rho i}/d) \tag{7}$$

3 Buckingham's Π Theorem Applied to Monocular SLAM

Buckingham's Π Theorem [1] is a key theorem in Dimensional Analysis. It states that physical laws are independent of units. Given a dimensionally correct equation involving n quantities of different kinds: $f(X_1, X_2, X_3, ..., X_n) = 0$ the existing relationship between the variables can be expressed also as:
$F(\Pi_1, \Pi_2, \Pi_3, ..., \Pi_{n-k}) = 0$ where Π_i is a reduced set of $n - k$ independent

dimensionless groups of variables, and k the number of independent dimensions that appear in the problem.

The monocular estimation process can be expressed as a function:

$$\left(\mathbf{r}^{WC}, \mathbf{q}^{WC}, \mathbf{v}^{W}, \omega^{W}, \mathbf{y}_1, \ldots, \mathbf{y}_n\right)^{\top} = \mathbf{f}(\sigma_a, \sigma_\alpha, \sigma_z, \mathbf{z}, \Delta t, \rho_0, \sigma_{\rho_0}, \sigma_{v0}, \sigma_{\omega 0}) \ , \quad (8)$$

where vector $\mathbf{z}$ stacks all the image measurements along the image sequence. Table 1 summarizes all the variables involved in monocular SLAM estimation and and their units.

$\mathbf{r}$	$\mathbf{q}$	$\mathbf{v}$, σ_{v0}	ω, $\sigma_{\omega 0}$		$\mathbf{z}$,$\sigma_{\mathbf{z}}$		a^W, σ_a	α^C,σ_α	x_i, y_i, z_i	θ_i, ϕ_i	$\rho_{\mathbf{i}}$, σ_{ρ_0}
l	1	lt^{-1}	t^{-1}	l^{-1}	lt^{-1}	t^{-1}	l^{-1}	1	1	lt^{-2}	t^{-2}

Table 1. Dimensionless parameters and the corresponding variables involved

Based on the equation above, dimensionless groups must be chosen. The parameters ρ_0 and Δt are the parameters of the two dimensions involved (length and time) chosen to form the dimensionless groups. (Table 2).

$\mathbf{\Pi_r}$	$\mathbf{\Pi_q}$	$\mathbf{\Pi_v}$	$\mathbf{\Pi}_\omega$	$\mathit{\Pi}_{\rho_i}$	$\mathit{\Pi}_{\sigma_{v0}}$	$\mathit{\Pi}_{\sigma_{\omega 0}}$	$\mathit{\Pi}_{\sigma_{\rho 0}}$	$\mathbf{\Pi_z}$	$\mathit{\Pi}_{\sigma z}$	$\mathit{\Pi}_{\sigma a}$	$\mathit{\Pi}_{\sigma \alpha}$
$\mathbf{r}\rho_0$	$\mathbf{q}$	$\mathbf{v}\rho_0\Delta t$	$\omega\Delta t$	$\frac{\rho_i}{\rho_0}$	$\sigma_{v0}\rho_0\Delta t$	$\sigma_{\omega 0}\Delta t$	$\frac{\sigma_{\rho 0}}{\rho_0}$	z	σz	$\sigma_a \rho_0 \Delta t^2$	$\sigma_\alpha \rho_0 \Delta t^2$

Table 2. Dimensionless numbers and the corresponding involved variables

4 Dimensionless Monocular SLAM Model

The state vector is composed of dimensionless parameters defining camera location, rotation and velocities, and the map features:

$$\mathbf{x_v} = (\mathbf{\Pi_r}, \mathbf{q}, \mathbf{\Pi_v}, \mathbf{\Pi}_\omega)^{\top} \quad \mathbf{\Pi_{y_i}} = (\mathbf{\Pi_{r_i}}, \theta_i, \phi_i, \mathit{\Pi}_{\rho_i})^{\top} \quad (9)$$

The dimensionless state update equation is:

$$\mathbf{f_v} = \begin{pmatrix} \mathbf{\Pi_r}_{k+1}^{WC} \\ \mathbf{q}_{k+1}^{WC} \\ \mathbf{\Pi_v}_{k+1}^{WC} \\ \mathbf{\Pi}_{\omega\, k+1}^{WC} \end{pmatrix} = \begin{pmatrix} \mathbf{\Pi_r}_k^{WC} + \mathbf{\Pi_v}_k^{WC} + \mathbf{\Pi_a}_k^{WC} \\ \mathbf{q}_k^{WC} \times \mathbf{q}(\mathbf{\Pi}_{\omega k}^{WC} + \mathbf{\Pi}_{\alpha k}^{WC}) \\ \mathbf{\Pi_v}_k^{WC} + \mathbf{\Pi_a}_k^{WC} \\ \mathbf{\Pi}_{\omega k}^{WC} + \mathbf{\Pi}_{\alpha k}^{WC} \end{pmatrix} \quad (10)$$

Next the monocular camera measurement equation is detailed. First, features coded in inverse depth must be converted to 3D points in the world reference:

$$\mathbf{\Pi}_h^W = \mathbf{\Pi_{r_i}} + \mathit{\Pi}_{\rho_i} \mathbf{m}(\theta_i, \phi_i) \ , \quad (11)$$

where $\mathbf{m}(\theta_i, \phi_i)$ is the unit vector defined by the pair of azimuth-elevation angles.

These world-referenced 3D points are converted to the camera frame:

$$\mathbf{\Pi}_h^C = \mathbf{R}^{CW}(\mathbf{\Pi}_h^W - \mathbf{\Pi_r}) \ , \quad (12)$$

and then are projected into the camera using the pinhole model:

$$\upsilon = \frac{\mathbf{\Pi}_h^C|_x}{\mathbf{\Pi}_h^C|_z} \quad \nu = \frac{\mathbf{\Pi}_h^C|_y}{\mathbf{\Pi}_h^C|_z} \tag{13}$$

Finally, camera calibration including radial distortion is applied to obtain pixel coordinates from angular coordinates.

There camera measurements clearly do not involve the size of the scene. If the metric parameter d has to be estimated, other types of measurements must be made. For instance, the equation that gives the distance between two points:

$$\mathcal{D}(\mathbf{P_1}, \mathbf{P_2}) = d\sqrt{(\mathbf{\Pi_{y_2}}|_x - \mathbf{\Pi_{y_1}}|_x)^2 + (\mathbf{\Pi_{y_2}}|_y - \mathbf{\Pi_{y_1}}|_y)^2 + (\mathbf{\Pi_{y_2}}|_z - \mathbf{\Pi_{y_1}}|_z)^2} \tag{14}$$

5 Image interpretation of dimensionless parameters and image filter tuning

The most representative of the dimensionless parameters can be seen in Figure 1. Their geometrical interpretation as camera angles is detailed here.

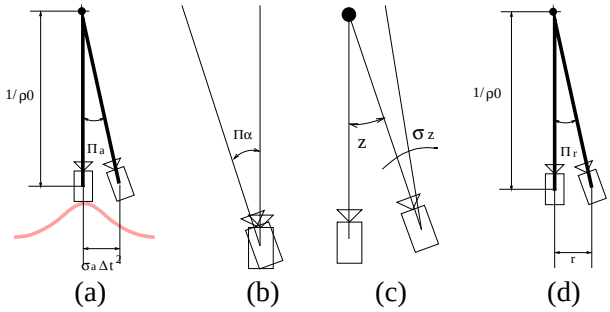

Fig. 1. Dimensionless monocular SLAM parameters.

Figure 1(a) shows the dimensionless parameter $\mathit{\Pi}_{\sigma a}$. The product $\sigma_a \Delta t^2$ represents the effect of the acceleration noise on the camera location. This value divided by $1/\rho_0$ gives the angle represented in the figure. This angle can be seen as the parallax allowed to a feature at depth $1/\rho_0$ due to camera acceleration.

The camera angular acceleration covariance in Figure 1(b) can clearly be interpreted as an angle between frames, and can be mapped to image pixels. Image measurements and image noise, in Figure 1(c) are directly measured in the image, so they are already dimensionless angles.

The translation estimate, $\mathit{\Pi}_r$ (in fig 1(d)), can also be seen as the angle defined by the translation between frames and the initial inverse depth.

As a consequence of this interpretation, EKF tuning is greatly simplified. Image values, observable in an image sequence, replace non-observable 3D real world values. Tuning parameters are related to image motion and no assumptions on the 3D scene are done.

6 Real Image Results

Real image experiments without adding any target to the scene has been performed. The first one shows how the scale of the scene in usual monocular SLAM depends on the prior knowledge of the scene. The second one illustrates the use of image tuning and the reduction in the number of tuning parameters. In the third experiment, the same image tuning is used in two different sequences which have different metric qualities but lead to the same image motion. All of the sequences have been recorded with a IEEE 1394 320×240 monochrome camera at 30 fps. A wide angle lens is used.

6.1 Dependence of scene scale on a priori parameters

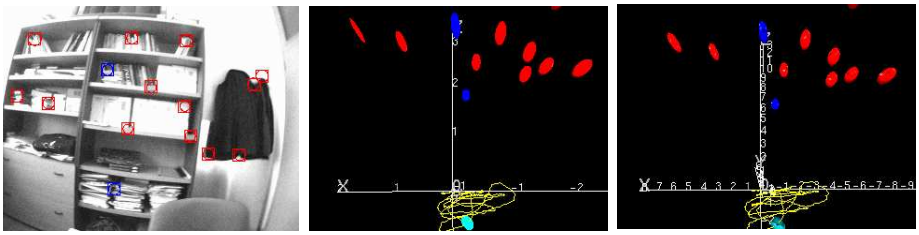

Fig. 2. Left: sample. Centre: EKF SLAM estimation result $\rho_0 = 0.5m^{-1}$. Right: EKF SLAM estimation result $\rho_0 = 0.1m^{-1}$.Feature uncertainty in red and blue, the camera uncertainty in cyan and the camera trajectory in yellow.

The same sequence was processed with the dimensional EKF SLAM algorithm varying the ρ_0 parameter. Figure 2 shows the estimation for $\rho_0 = 0.5m^{-1}$ and $\rho_0 = 0.1m^{-1}$. Notice that the estimated depth of the scene (the distance between the camera and the points in the bookcase) tends to be at the depth prior ($2m$ and $10m$). The two estimated scenes have the same form, the difference is just the scale of the axis. If $\mathbf{\Pi_r^{WC}} = \rho_0 \mathbf{r^{WC}}$ and $\mathbf{\Pi_{y_i}}$ were estimated using the dimensionless monocular SLAM proposed, these two experiments would be normalized into one, in which normalized depth tends to be at *dimensionless* '1',

6.2 Image tuning in a pure rotation sequence

This sequence is a pure camera rotation in a hallway. Dimensional monocular SLAM should have been tuned with real camera accelerations and depth priors. As these values are not observable, they need to be assumed. Dimensionless monocular SLAM is tuned directly with image values.

Two experiments with the same $\Pi_{\sigma a} = 0$, $\Pi_{\sigma z} = 1pxls$ values but different tuning in $\Pi_{\sigma\alpha}$: a)$\Pi_{\sigma\alpha} = 2pxls$, and b)$\Pi_{\sigma\alpha} = 4pxls$ has been performed (Fig.3.) Because of the image tuning, their effect can be directly seen in the 95% image search regions size for the map features.

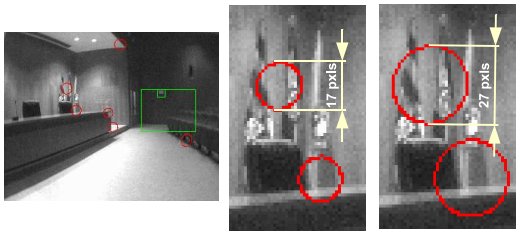

Fig. 3. Pure rotation image search regions. Left: sample image. Centre: $\Pi_\alpha = 2pxls$. Right:$\Pi_\alpha = 4pxls$.

It is important to notice that, in the previous paragraph, neither 3D scene assumptions nor time between frames Δt are needed in the filter. The tuned values are the allowed image motion between frames due to camera linear and angular acceleration and image noise.

6.3 The same image tuning for different sequences

Two translational sequences have been recorded walking along a corridor and looking at the wall. In the first one, the distance from the wall was 2.5 metres. In the second, the distance from the wall was twice (5 metres), the distance walked along the corridor the same, and the walking velocity was double (therefore, the number of frames of the second sequence is half the first one). Although they are two different experiments, the image motion in both sequences is the same, and dimensionless monocular SLAM has to be tuned with same values. In this experiment, these values were: $\sigma_z = 1pxl$, $\sigma_a = 2pxl$ and $\sigma_\alpha = 2pxl$. Notice again the simplicity of image tuning compared with 3D tuning, in which you have to imagine the depth prior and the 3D accelerations, unobservable with a single camera. Figure 4 shows the results of both estimations.

The dimensionless estimated translation can be interpreted as the translation in units of the initial depth prior. As the wall is twice as far in the second sequence, the second sequence's estimated translation is half. It can also be noticed that, as the normalized translation is smaller in the second experiment, the normalized 3D point positions are estimated with less accuracy and have larger uncertainty regions.

7 Conclusions

Up-to-scale results from real-time, EKF based monocular SLAM without an initialisation target are presented. As no known points are included in the estimation, the real size of the scene cannot be recovered. Nevertheless, a scaled estimation is obtained, its size depending on priors introduced to the filter.

In order to represent the non-observability of the real size of the scene, a new monocular SLAM parameterisation is presented. This approach separates

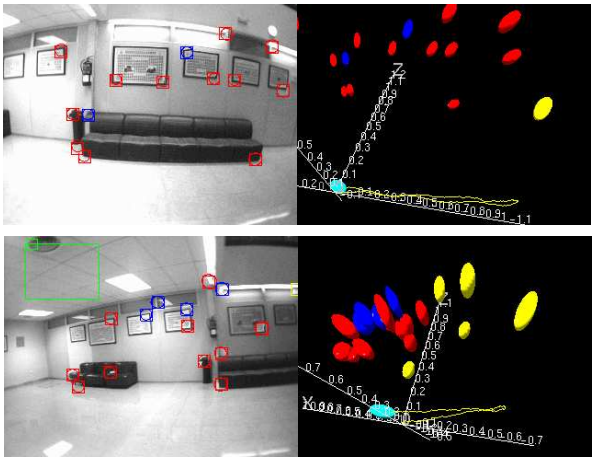

Fig. 4. Two equivalent sequences. First and last images and 3D estimated geometry

the geometric problem of estimating a point map and camera motion up to scale from the unobservable real size of the map and motion. A parameter that codes the real size of the scene is added to the state vector, but single-camera measurements do not involve this value. As a consequence, its value cannot be estimated with single camera measurements.

Buckingham's theorem was used to build the dimensionless state vector in this new EKF approach. A geometrical interpretation of the dimensionless parameters as angles allows a simplified tunning of the filter: the number of tuning parameters is reduced and no 3D assumptions of the scene are made.

References

1. E. Buckingham. On physically similar systems; illustrations of the use of dimensional equations. *Phys. Rev.*, 4(4):345–376, Oct 1914.
2. J. A. Castellanos and J. D. Tardós. *Mobile Robot Localization and Map Building: A Multisensor Fusion Approach*. Kluwer Academic Publishers, Boston. USA, 1999.
3. A. J. Davison. Real-time simultaneous localization and mapping with a single camera. In *Proc. International Conference on Computer Vision*, 2003.
4. E. Eade and T. Drummond. Scalable monocular SLAM. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
5. R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.
6. J. Montiel, J. Civera, and A. Davison. Unified inverse depth parametrization for monocular slam. In *Proceedings of Robotics: Science and Systems*, Philadelphia, USA, August 2006.
7. R. Smith and P. Cheeseman. On the representation and estimation of spatial uncertainty. *Intl. Journal of Robotics Research*, 5(4):56–68, 1986.