Proceedings of the 2006 IEEE International Conference on Robotics and Automation
Orlando, Florida - May 2006

Active Control for Single Camera SLAM

Teresa Vidal-Calleja
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens Artigas 4-6, Barcelona 08028, Spain.
tvidal@iri.upc.edu

Andrew J. Davison
Department of Computing, Imperial College London
180 Queen's Gate, London SW7 2AZ, UK.
ajd@doc.ic.ac.uk

Juan Andrade-Cetto
Centre de Visiò per Computador, UAB
Edifici O, Campus UAB, Bellaterra 08193, Spain.
cetto@cvc.uab.es

David W. Murray
Robotics Research Group, Department of Engineering Science
University of Oxford, Oxford OX1 3PJ, UK.
dwm@robots.ox.ac.uk

***Abstract*—**

In this paper we consider a single hand-held camera performing SLAM at video rate with generic 6DOF motion. The aim is to optimise both the localisation of the sensor and building of the feature map by computing the most appropriate control actions or movements. The actions belong to a discrete set (e.g. go forward, go left, go up, turn right, etc), and are chosen so as to maximise the mutual information gain between posterior states and measurements. Maximising the mutual information helps the camera avoid making ill-conditioned measurements appropriate to bearing-only SLAM. Moreover, orientation changes are determined by maximising the trace of the Fisher Information Matrix. In this way, we allow the camera to continue looking at those landmarks with large uncertainty, but from better-posed directions. Various position and gaze control strategies are first tested in a simulated environment, and then validated in a video-rate implementation. Given that our system is capable of producing motion commands for a real-time 6DOF visual SLAM, it could be used with any type of mobile platform, without the need of other sensors.

I. Introduction

Impressive advances in 2D and, more recently, 3D simultaneous localisation and mapping (SLAM) for mobile robots have been made over the last 15 years, largely using sonar and laser range sensing [1]–[5]. Most recently, there has been considerable interest in solving the SLAM problem using visual sensing, both in order to obtain more accurate 3D representations of the environment and to exploit its richer potential for scene representation [6], [7]. In this communication, we consider the problem of SLAM with a single camera carried by a human, and how to implement control strategies in this context. In that sense, this work is different from other control work because we can only give a human quite approximate, low frequency, easy to understand commands like 'left', 'right', 'stay'.

One of the first active vision-based SLAM approaches used feature correspondences from stereo image pairs [6]. The computational burden for the accurate detection and matching of image pairs motivated the use of active visual sensing for landmark selection in sparse feature maps. Their work is different to ours because they only control orientations of the stereo head, and we are now talking about actually controlling translation as well. Other reported techniques to visual SLAM — although with no control — include the use of SIFT features, and matching over a trinocular rig [7]. More elegantly and economically, feature locations can also be computed by tracking landmarks over multiple views from only one camera, a process referred to a 'bearing-only SLAM'.

One key issue in bearing-only SLAM is the initialisation of feature locations. In [8] for example, the initial estimation of a landmark's location is achieved by sampling hypotheses of a 1D particle distribution along the line of sight. Another technique consists of using sums of Gaussian distributions to parameterise 3D feature locations over a delayed state representation [9].

When the sensor capabilities in SLAM are limited, camera motion plays an important role in the quality of reconstruction obtained. Driving the sensor to the locations that maximise the expected information gain from acquiring an observation at that location has been a common strategy [10]–[12]. However, Sim has showed that maximising the expected information gain leads to ill-conditioned filter updates in the bearing-only SLAM [13]. In [14], Bryson *et al.* present simulated results of the effect different vehicle actions have with respect to the entropic mutual information gain. The analysis is performed for a 6DOF aerial vehicle equipped with two cameras and an inertial sensor, for which landmark range, azimuth, and elevation readings are simulated, and data association is known.

In this paper we are interested in the video-rate estimation *and control* of a single camera's motion, moving rapidly with 6DOF in 3D in normal human environments, mapping visual features with minimal prior information about motion dynamics. Our aim is to localise the sensor and build a feature map by computing the appropriate control actions in order to improve overall system estimation.

However, insisting on video-rate performance using modest hardware imposes severe restrictions on the volume of

This research is supported by the Spanish Ministry of Education and Science under projects DPI 2004-5414, TIC 2003-09291, and the EU PACO-PLUS project FP6-2004-IST-4-27657 to TVC and JAC, and by the UK Engineering and Physical Research Council under an Advanced Research Fellowship Grant GR/N03266 to AJD, and by Grant GR/S97774 to DWM.

0-7803-9505-0/06/$20.00 ©2006 IEEE

computation that can take place in each 33ms time step. Re-estimation must take place of course, but making strictly optimal camera movements would require in addition the computation of the derivatives of a well-chosen performance metric with respect to the inputs [15]. Such a computation remains unfeasible for a 6DOF highly nonlinear system model. Besides, human actions can only be approximate, and at low frequency. So, instead of computing the optimal motion command, we decide only upon a small set of choices.

Actions belong to a discrete set (eg. go forward, go left, go up, turn right, etc.), and the particular movement chosen is the one that maximises the mutual information gain between posterior states and measurements. Using entropy for exploration only makes sense if we can be certain that uncertainty is reduced as landmarks are being discovered. To that, one must have an idea first of the shape of the space to be mapped, and filling it with randomly placed features with large uncertainty [14]. Maximising the mutual information aims at reducing the overall state uncertainty, and helps the camera move away from making repeated ill-conditioned measurements. Orientation changes are determined by maximising the trace of the Fisher Information Matrix. In this way, we allow the camera still to look at those landmarks with large uncertainty, but from better-posed directions.

The remainder of the paper is ordered as follows. First we briefly describe the system and the estimation scheme. Then the metrics used as cost functions to choose the appropriate actions are explained; and our control strategy is illustrated through simulations. Lastly, we present the results of real-time experiments with a hand-held wide-angle camera, where a GUI feeds-back motion commands to the user.

II. 6 DOF Bearing-Only SLAM

A. Unconstrained Camera Motion

It is assumed that the camera could be attached to any mobile platform — in our case the hand — and is free to move in any direction in $\mathbb{R}^3 \times SO(3)$. We adopt a smooth unconstrained constant-velocity motion model, its translational and rotational altered only by zero-mean, normally distributed accelerations and staying the same on average. The Gaussian acceleration assumption means that large impulsive changes of direction are unlikely. The camera motion prediction model is

$$\mathbf{x}_{v(k+1|k)} = \begin{bmatrix} \mathbf{p}_{(k+1|k)} \\ \mathbf{q}_{(k+1|k)} \\ \mathbf{v}_{(k+1|k)} \\ \boldsymbol{\omega}_{(k+1|k)} \end{bmatrix} = \begin{bmatrix} \mathbf{p}_{(k|k)} + (\mathbf{v}_{(k|k)} + \mathbf{a}_{(k)}\Delta t)\Delta t \\ \boldsymbol{\mathcal{Q}}\mathbf{q}_{(k|k)} \\ \mathbf{v}_{(k|k)} + \mathbf{a}_{(k)}\Delta t \\ \boldsymbol{\omega}_{(k|k)} + \boldsymbol{\alpha}_{(k)}\Delta t \end{bmatrix},$$

with $\mathbf{p} = [x, y, z]^\top$ and $\mathbf{q} = [q_0, q_1, q_2, q_3]^\top$ denoting the camera pose (three states for position and four for orientation using a unit norm quaternion representation), and $\mathbf{v} = [v_x, v_y, v_z]^\top$ and $\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]^\top$ denoting the linear and angular velocities, respectively. The subscripts $(k|k)$ and $(k+1|k)$ denote the posterior at time k and the prior (before integrating measurements) at $k+1$. The input to the system is the acceleration vector $\mathbf{u} = [\mathbf{a}^\top, \boldsymbol{\alpha}^\top]^\top = [a_x, a_y, a_z, \alpha_x, \alpha_y, \alpha_z]^\top$.

An Extended Kalman Filter propagates the camera pose and velocity estimates, as well as feature estimates. A state that includes the features $\mathbf{y}$ is made of $\mathbf{x} = [\mathbf{x}_v^\top, \mathbf{y}^\top]^\top$. The model $\boldsymbol{\mathcal{Q}}$ for the prediction of change in orientation is inspired by [16] and is detailed in the Appendix. The redundancy in the quaternion representation is removed by a $||\mathbf{q}|| = 1$ normalisation at each update, accompanied by the corresponding Jacobian modification.

B. Feature Extraction

In this work we are interested in mapping the 3-D coordinates of salient point features from images, and need to do so at video-rate. As in previous work, we use the Shi-Tomasi saliency operator, and match correspondences in subsequent frames using normalised sum-of-squared differences [6], [8]. Although more robust detectors such as SIFT have become widely popular for their ability to find and match features with higher degree of uniqueness, they come at the expense of heavier computational load.

Image projection is modelled using a full perspective wide angle camera. The position of a 3D scene point $\mathbf{y}_i$ is transformed into the camera frame as $\mathbf{y}_i^c = [x^c, y^c, z^c]^\top = \boldsymbol{\mathcal{R}}^\top(\mathbf{y}_i - \mathbf{p})$, with $\boldsymbol{\mathcal{R}}$ the rotation matrix equivalent of $\mathbf{q}$. The point's projection onto the image plane is

$$\mathbf{h}_i = \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u_0 - u_c/\sqrt{d} \\ v_0 - v_c/\sqrt{d} \end{bmatrix}, \quad (1)$$

where $u_c = fk_u x^c/z^c$, $v_c = fk_v y^c/z^c$, the radial distortion term is $d = 1 + K_d(u_c^2 + v_c^2)$, and the intrinsic calibration of the camera — focal distance f, principal point (u_0, v_0), pixel densities k_u and k_v, and radial distortion parameter K_d — are determined beforehand.

When an image feature is detected, its measurement must either be associated with an existing feature or be added as a new feature in the map. The location of the camera, along with the locations of the already mapped features, are used to predict feature position $\mathbf{h}_i$ using Eq. (1), and these estimates checked against the measurements using a nearest neighbour test. Feature search is constrained to 3σ elliptical regions around the image estimates as defined by the innovation covariance matrix $\mathbf{S}_i = \mathbf{H}_i\mathbf{P}_{k+1|k}\mathbf{H}_i^\top + \mathbf{R}$, with $\mathbf{H}_i$ the Jacobian of the sensor model with respect to the state, $\mathbf{P}_{k+1|k}$ the prior state covariance, and measurements $\mathbf{z}_i$ assumed corrupted by zero mean Gaussian noise with covariance $\mathbf{R}$.

C. Initialisation

Inserting a new feature to the map cannot be done immediately because the measurement model is non-invertible. Though bearing is recoverable from one measurement, 3D depth is not.

Several schemes have been reported [8], [9], [17], and we adopt the first of these. The initial measurement results in a semi-infinite line with Gaussian uncertainty in its parameters, starting at the estimated camera position and heading to infinity along the feature viewing direction. A 1D particle

distribution represents the likelihood of the 3D feature's position along this line. The line is projected as an epipolar line into subsequent images, but specifically it is the projection of the point particles and their uncertainly ellipses that provide the regions to be searched for a match, in turn producing likelihoods for Bayesian re-weighting of the depth distribution. A small number of steps is required to reduce to below a threshold the ratio of the standard deviation in depth to the depth estimate itself. At that time, the depth distribution is re-approximated as Gaussian and the feature is initialised as a 3D point $\mathbf{y}_i$ into the map.

III. Information Gain

This section first presents a metric for expected information gain as a result of performing a given action, and then develops an overall information conditioning strategy for the computation of orientations. The aim will be to move the camera in the direction that most reduces the uncertainty in the entire SLAM state, by using the information that should be *gained* from future, predicted, landmark observations were such a move to be made, but taking into account the information *lost* as a result of moving with uncertainty.

A. Mutual Information Gain

We adopt entropy as a measure of uncertainty; that is, as a measure of how much randomness there is in our state estimate. Entropy is defined as $H(X) = -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x})$, which, for our case where $p(\mathbf{x})$ is a n-variate Gaussian distribution, reduces to $H(X) = \frac{1}{2}\log((2\pi)^n|\mathbf{P}|)$.

Now consider the following two random vectors: the state prior $\mathbf{x}_{k+1|k}$, and the prediction of measurement i, $\mathbf{z}_{i,k+1|k}$. We want to choose the action that maximises the mutual information between the two. The mutual information is defined as the relative entropy between the joint distribution $p(\mathbf{x}, \mathbf{z}_i)$, and the marginals $p(\mathbf{x})$ and $p(\mathbf{z}_i)$.

$$\begin{aligned} I(X;Z) &= \sum_{\mathbf{x}\in X, \mathbf{z}_i \in Z} p(\mathbf{x}, \mathbf{z}_i) \log \frac{p(\mathbf{x}, \mathbf{z}_i)}{p(\mathbf{x})p(\mathbf{z}_i)} \\ &= H(X) + H(Z) - H(X,Z) \\ &= H(X) - H(X|Z)\,, \end{aligned}$$

which, for our Gaussian multivariate case, evaluates to

$$\begin{aligned} I(X;Z) &= \frac{1}{2}\log\left(\frac{|\mathbf{P}_{\mathbf{x}}|}{|\mathbf{P}_{\mathbf{x}} - \mathbf{P}_{\mathbf{xz}}\mathbf{P}_{\mathbf{z}}^{-1}\mathbf{P}_{\mathbf{xz}}^{\top}|}\right) \\ &= \frac{1}{2}\log\left(\frac{|\mathbf{P}_{k+1|k}|}{|\mathbf{P}_{k+1|k} - \mathbf{P}_{k+1|k}\mathbf{H}_i^{\top}\mathbf{S}_i^{-1}\mathbf{H}_i\mathbf{P}_{k+1|k}^{\top}|}\right) \\ &= \frac{1}{2}\left(\log|\mathbf{P}_{k+1|k}| - \log|\mathbf{P}_{k+1|k+1}|\right)\,. \end{aligned}$$

Thus, in choosing a maximally mutually informative motion command, we are maximising the difference between prior and posterior entropies [18]. In other words, we are choosing the motion command that most reduces the uncertainty of $\mathbf{x}$ due to the knowledge of $\mathbf{z}$ as a result of a particular action. Figure 1 shows the directions maximising the mutual information for a simple 2DOF camera and 3 landmarks.

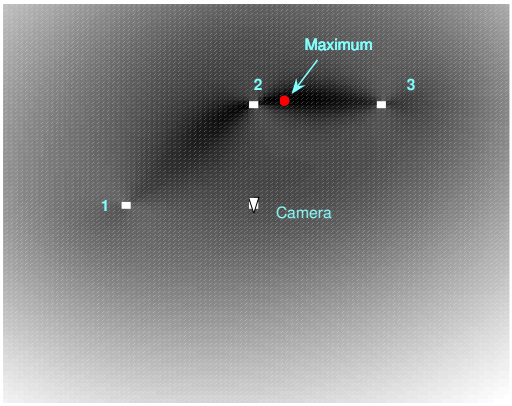


Fig. 1. Maximisation of mutual information for the evaluation of motion commands. A simple 2DOF camera is located at the centre of the plot, and a decision where to move must be taken as a function of the pose and landmark states, and the expected measurements. Three landmarks are located to its left, front, and right-front. Moving to the location in between landmarks 2 and 3 maximises the mutual information between the SLAM prior and the measurements for this particular example.

Note that the use of mutual information only makes sense prior to reaching full correlation. In SLAM, $|\mathbf{P}_{k|k}|$ tends asymptotically to zero, point at which the map becomes fully correlated and there is nothing else the camera can do to improve the estimates of the features. From then on, entropy can still be used to decide what actions to take to reduce the camera's own uncertainty, and this can be done just by replacing $\mathbf{x}$ with $\mathbf{x}_v$ from the above discussion.

B. Fisher Information for Gaze Direction

Measurements in the bearing-only SLAM case are ill-posed for motions along the principal axis, when points are close to the principal axis and there is little perspective distortion. Motion commands based on the maximisation of the mutual information metric drive the camera away from those configurations, that is, perpendicular to the principal axis. However, we still want the camera to look at those landmarks with large uncertainty so as to reduce their covariance when seen from different locations. To do that, we incorporate another information metric to control the direction of gaze. From a set of possible orientation changes, we propose choosing that which maximises the trace of the Fisher Information Matrix. In this way we will be choosing the best direction to look at, in the sense that it is the one that is most informative, but from a different position than the ill-posed one. Under the Gaussian assumption for sensor and platform noises, the minimisation of the least squares criteria (the KF) is equivalent to the maximisation of a likelihood function $\Lambda(\mathbf{x})$ given the set of observations Z^k, that is, the maximisation of the joint pdf of the entire history of observations, $\Lambda(\mathbf{x}) = \prod_{i=1}^{k} p(\mathbf{z}_i|\mathbf{x}, Z^{i-1})$.

The Total Fisher Information Matrix, a quantification of the maximum existing information in the observations

about the state, is defined in [19] as the expectation $\mathbf{J} = E\left[(\nabla \log \Lambda(\mathbf{x}))(\nabla \log \Lambda(\mathbf{x}))^{\top}\right]$, which here evaluates to $\mathbf{J} = \sum \mathbf{H}^{\top}\mathbf{S}^{-1}\mathbf{H}$.

The information for the reconstruction of the state contributed by the set of measurements at each iteration is contained in $\mathbf{H}^{\top}\mathbf{S}^{-1}\mathbf{H}$. The eigenvalues λ_j of this contribution to $\mathbf{J}$ show which linear combinations of the states can be estimated with good accuracy and which will have large uncertainties from the coming measurements. It also shows which linear combinations of states are unobservable. When one dimension of $\mathbf{J}$ has a very small eigenvalue (information along the line of sight), the product is not a reliable measure of the elongation of the information hyperellipsoid, as it collapses the volume to zero. Our strategy is to look in the direction at which $\sum \lambda_j$ is maximum [20]. This is the viewing direction that will introduce the largest amount of information in one single measurement step.

Under a Fisher information motion strategy, maximally informative actions move the robot as close as possible to the landmarks under observation. We do not want to move towards them, but only to orient towards them. Our idea of using the Fisher Information is only to fixate our camera to those most uncertain landmarks, and use the change in entropy to select movement actions. This way, by using the mutual information metric, maximally informative actions would prevent the camera from producing ill-posed measurements. Note that an omnidirectional sensor would not require a strategy to direct fixation. In our case, as opposed to a mobile robot, translation and orientation changes are kinematically decoupled, for this reason, it makes sense to use different information measures in evaluating them.

IV. Control Strategy

In this Section we demonstrate in simulation how combining the strategies of effectively controlling translation by maximising mutual information thereafter controlling orientation by maximising the information available from the new position yields reliable active control of pose and velocity for a free moving camera, whilst building a map optimally.

A. Deciding Where to Go and Where to Look At

As noted earlier, the real-time requirements of the task preclude using an optimal control decision that takes into account all possible motion commands which is impracticable to compute, leading to an exponential growth because of the curse of dimensionality of long term action evaluation. Instead we evaluate our information metrics for a small set of actions carried out over a fixed amount of time, and choose the best action from those.

The set of possible actions is divided in two groups. Mutual information is evaluated for the translational actions `go_forward`, `go_backwards`, `go_right`, `go_left`, `go_up`, `go_down`, and `stay`; and Fisher information is maximised from the set of orientation commands `turn_right`, `turn_left`, and `stay`.

In our simulated setting, desired camera locations are predicted for the best action chosen, and a PD low-level control law is applied to ensure these locations are reached at the end of one second; at which point the motion metric is again evaluated to determine the next desired action. Orientations however, are evaluated at frame rate, leaving the system to freely rotate, governed only by the information maximisation strategy.

The simulation considers a fixed number of expected landmarks to be found, and both the Mutual Information and Fisher Information metrics are computed taking into account the corresponding full covariance matrices, including these unvisited landmarks, which have been initialised with large uncertainties. This is the only thing that prevents our control strategies from defaulting to homeostasis.

B. Simulation Results

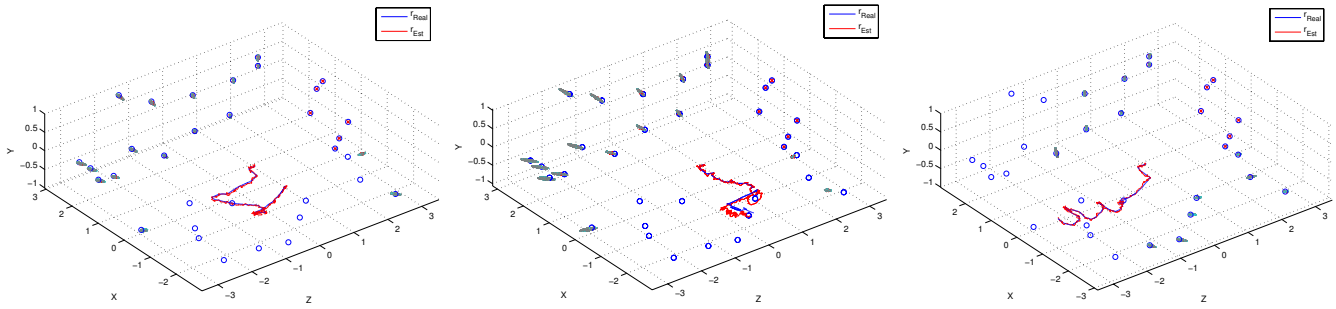
Figure 2 contains simulation results from our mutual information strategy for the computation of motion commands, and compares various orientation computation schemes. The simulated environment represents a room $6 \times 6 \times 2$ m^3 in size containing 33 randomly distributed point landmarks, out of which 6 are fiduciary points, to be used as global references [21].

The initial standard deviation in camera pose is 6-cm in the x and y directions, 4.6 cm in height z, and 45° in orientation, right after matching the fiduciary points, but before any motion takes place. Sensor standard deviation is set at 2 pixels, and data association is not known a priori. Instead, nearest neighbour χ-squared tests are computed to guarantee correct matching. New features are initialised once their ratio of depth estimate to depth standard deviation falls below a threshold of 0.3.

The plots show the results of actively moving a 6-DOF camera whilst building a map of 3D features. In all cases, each of the seven motion actions will produce a displacement of 30 cm in the corresponding direction. Our mutual information metric is evaluated at each of these positions. The action that maximises the metric is chosen, and the camera is controlled to reach that position in one second with a PD control law. Orientation changes are computed every 50 ms.

Three approaches were tested for the computation of gaze commands: (i) constant rotational velocity of 0.2 rad/sec, frames (a,d); (ii) maximisation of mutual information both for the position and orientation of the moving camera, frames (b,e); and (iii) maximisation of mutual information for position and maximisation of Fisher information for gaze, frames (c,f). The experiment shown in the plots lasted 35 seconds.

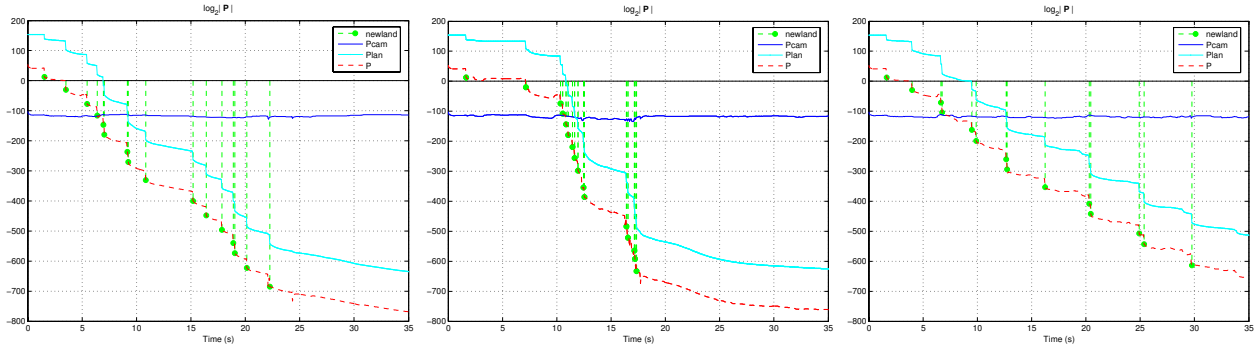
The constant rotational velocity and the mutual information strategies tend to insert landmarks into the map at a faster pace than the Fisher Information strategy. As can be seen in the error plots in Figure 3, this might not be always the best choice. It seems reasonable to let the system accurately locate the already seen landmarks before actively searching for new ones.



(a) Final Map by using Mutual Information for position and constant angular velocity.

(b) Final Map by using Mutual Information for position and orientation.

(c) Final Map by using Mutual Information for position and Fisher Information for orientation.



(d) Entropy for MI in position and constant angular velocity.

(e) Entropy for MI in position and orientation.

(f) Entropy for MI in position and FI in orientation.

Fig. 2. Trajectories with Final Maps and Entropy. ($\mathtt{r}_{\mathtt{Real}}$ and $\mathtt{r}_{\mathtt{Est}}$ are the real and estimated camera trajectories, the label `newland` and the green dots and dotted vertical lines represent the value of entropy at the instant when new landmarks are initialised. `Pcam`, `Plan`, and `P` indicate the camera, map, and overall entropies.

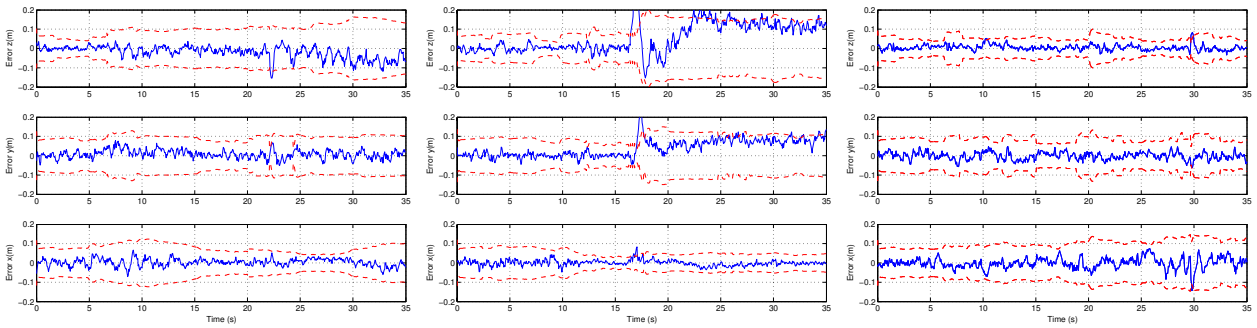
The third alternative, controlling camera orientation by maximising the Fisher Information entering into the filter, has the effect that it focuses on reducing the uncertainty of the already seen landmarks, instead of eagerly exploring the entire room for new landmarks. The reason is that landmarks that have been observed for a small period of time still have large depth uncertainty, and the Fisher Information metric is maximised when observations are directed towards them. The technique tends to close loops at a faster pace than the other two approaches, thus propagating correlations amongst landmarks and poses in a more efficient way. Additionally, by revisiting fiduciary points more often, orientations are much better estimated in this case.

Strategy (iii) needs more time to reduce entropy and takes more time to insert the same number of landmarks in the map. But, at the point at which the same number of landmarks is available it has lower entropy than the other two strategies (see for example in Figure 2, frames (d-f), that when the 14th landmark is added, the times are 19, 18, and 30 secs, and the entropies are -530, -550, and -610).

V. Experiments

This section presents an initial experimental result validating the maximisation of mutual information strategy for the control of a hand-held camera in a challenging 15fps visual SLAM application. Within a room, the camera starts approximately at rest with some known object in view to act as a starting point and provide a metric scale to the proceedings. The camera moves, translating and rotating freely in 3D, according to the instruction provided in a graphical user interface, and executed by the user, within a room or a restricted volume, such that various parts of the unknown environment come into view. The aim is to estimate and control the full camera pose continuously during arbitrarily long periods of movement. This involves accurately mapping (estimating the locations of) a sparse set of features in the environment.

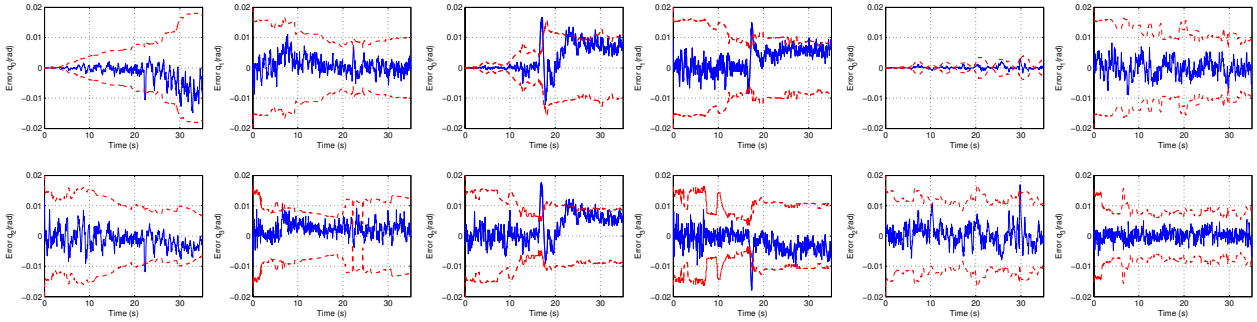
Given that the control loop is being closed by the human operator, only displacement commands are computed. Gaze control is left to the user. Furthermore, the mutual information measure requires evaluating the determinant of the full covariance matrix at each iteration. Because of the complexity of this operation, single motion predictions are evaluated one frame at a time. It is only until the 15th frame in the sequence that all mutual information measures are compared, and a desired action is displayed on screen. That is, the user is presented with motion directions to obey every second. Note also, that in computing the mutual information measure, only the camera position and map parts of the covariance matrix are used, leaving out the gaze and velocity parts of the matrix. Finally, to keep it running in real-time, the resulting application must be designed for sparse mapping. That is, with the computing capabilities of an off-the-shelf system, our current application



(a) Position error when using MI for position and constant angular velocity.

(b) Position error when using MI for position and orientation.

(c) Position error when using MI for position and FI for orientation.



(d) Orientation error when using MI for position and constant angular velocity.

(e) Orientation error when using MI for position and orientation.

(f) Orientation error when using MI for position and FI for orientation.

Fig. 3. Estimation errors for camera position and orientation and their corresponding 2σ variance bounds. Position errors are plotted as x, y, and z distances to the real camera location in meters, and orientation errors are plotted as quaternions.

is limited to less than 50 landmarks.

Figure 4 shows the graphical user interface. The top part of the figure contains a 3D plot of the camera and the landmarks mapped, while the bottom part shows the information being displayed to the user superimposed on the camera view. Figure 5 contains a plot of the decrease in the various entropies for the map being built, and the list of actions chosen as shown to the user during the first minute.

Worth noticing is that in the real-time implementation, the system prompts the user for repeated up-down movements, as well as left-right commands. This can be explained as if after initialising new features, the system repeatedly asks for motions perpendicular to the line of sight to best reduce their uncertainty. Also, closing loops has an interesting effect in the reduction of entropy, as can be seen around the 1500th frame on Fig. 5-a.

VI. Conclusion

In conclusion, we have shown plausible motion strategies in a video-rate visual SLAM application. On the one hand, by choosing a maximal mutually informative motion command, we are maximising the difference between prior and posterior SLAM entropies, resulting in the motion command that mostly reduces the uncertainty of $\mathbf{x}$ due to the knowledge of $\mathbf{z}$. Alternatively, by controlling gaze maximising the information about the measurements, we get a system that prioritises in accurately locating the already seen landmarks before actively searching for new ones.

Our method is validated in a video-rate hand-held visual SLAM implementation. Given that our system is capable of producing motion commands for a real-time 6DOF visual SLAM, it is sufficiently general to be incorporated into any type of mobile platform, without the need of other sensors.

A possible weakness of this information-based approach is that it estimates the utility of measurements assuming that our models are correct. Model discrepancies, and effects of linearisation in the computation of our estimation and control commands might lead to undesirable results.

Appendix

The orientation of the camera frame, and its rate of change, are related to the angular velocity by the quaternion multiplication $\boldsymbol{\Omega} = 2\dot{\mathbf{q}}\mathbf{q}^*$, with $\boldsymbol{\Omega} = [0, \omega_x, \omega_y, \omega_z]^\top$, the angular velocity vector expressed in quaternion form, and $\mathbf{q}^*$ is the orientation quaternion conjugate. Or equivalently, by $\dot{\mathbf{q}} = \frac{1}{2}\mathbf{M}\mathbf{q} \approx \frac{\mathbf{q}_{(k+1)} - \mathbf{q}_{(k)}}{\Delta t}$, with

$$\mathbf{M} = \begin{bmatrix} 0 & -\omega_x & -\omega_y & -\omega_z \\ \omega_x & 0 & -\omega_z & \omega_y \\ \omega_y & \omega_z & 0 & -\omega_x \\ \omega_z & -\omega_y & \omega_x & 0 \end{bmatrix}.$$

Solving for $\mathbf{q}_{(k+1)}$ in the above approximation when $\boldsymbol{\omega}$ is constant, our smooth motion model for the prediction of change

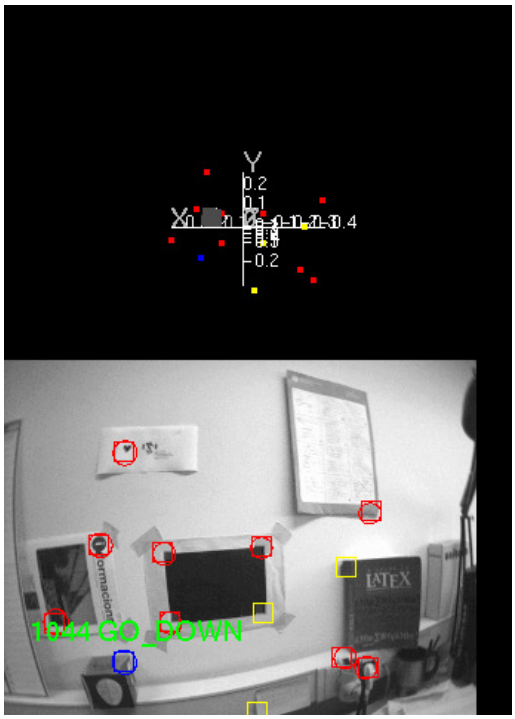
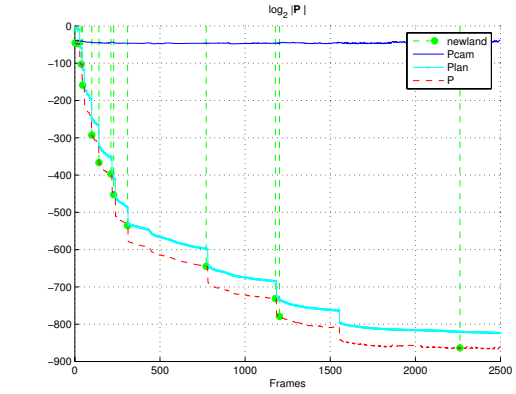
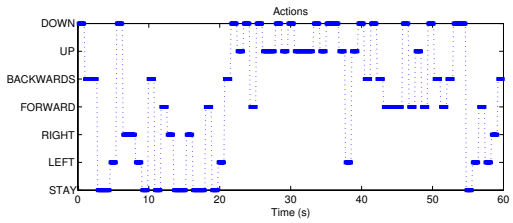


Fig. 4. Feature map and camera view as shown in the Graphical User Interface (844th frame).



(a) Camera, Map, and Total Entropies.



(b) Actions for the first minute

Fig. 5. Real-time Active Vision SLAM.

in orientation becomes $\mathbf{q}_{k+1} = \boldsymbol{\mathcal{Q}}\mathbf{q}_k$ with the quaternion transition matrix

$$\boldsymbol{\mathcal{Q}} = \cos\left(\frac{\Delta t\|\boldsymbol{\Omega}\|}{2}\right)\mathbf{I} + \frac{2}{\|\boldsymbol{\Omega}\|}\sin\left(\frac{\Delta t\|\boldsymbol{\Omega}\|}{2}\right)\mathbf{M}\,.$$

Note that when computing the quaternion propagation, the angular velocities are to be evaluated at $(k+1|k)$, i.e., including the angular acceleration term.

REFERENCES

[1] R. C. Smith and P. Cheeseman, "On the representation and estimation of spatial uncertainty," *Int. J. Robot. Res.*, vol. 5, no. 4, pp. 56–68, 1986.

[2] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localization and map building (SLAM) problem," *IEEE Trans. Robot. Automat.*, vol. 17, no. 3, pp. 229–241, Jun. 2001.

[3] F. Lu and E. Milios, "Globally consistent range scan alignment for environment mapping," *Auton. Robot.*, vol. 4, no. 4, pp. 333–349, 1997.

[4] U. Frese, P. Larsson, and T. Duckett, "A multigrid algorithm for simultaneous localization and mapping," *IEEE Trans. Robot.*, vol. 21, no. 2, pp. 1–12, 2005.

[5] S. Thrun, Y. Liu, D. Koller, A. Y. Ng, Z. Ghahramani, and H. Durrant-Whyte, "Simultaneous localization and mapping with sparse extended information filters," *Int. J. Robot. Res.*, vol. 23, no. 7-8, pp. 693–716, Jul. 2004.

[6] A. J. Davison and D. W. Murray, "Simultaneous localisation and map-building uisng active vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 7, pp. 865–880, Jul. 2002.

[7] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *Int. J. Robot. Res.*, vol. 21, no. 8, pp. 735–758, Aug. 2002.

[8] A. Davison, W. Mayol, and D. Murray, "Real-time localisation and mapping with wearable active vision," in *Proc. IEEE Int. Sym. Mixed and Augmented Reality*, Tokyo, Oct. 2003.

[9] J. Sola, A. Monin, M. Devy, and T. Lemaire, "Undelayed initialization in bearing only SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Edmonton, Aug. 2005.

[10] P. Whaite and F. Ferrie, "Autonomous exploration: Driven by uncertainty," in *Proc. 9th IEEE Conf. Comput. Vision Pattern Recog.*, Seattle, Jun. 1994, pp. 339–346.

[11] H. Feder, J. Leonard, and C. Smith, "Adaptive mobile robot navigation and mapping," *Int. J. Robot. Res.*, vol. 18, pp. 650–668, 1999.

[12] F. Bourgault, A. Makarenko, S. Williams, and B. Grocholsky, "Information based adaptative robotic exploration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Lausanne, Oct. 2002.

[13] R. Sim, "Stable exploration for bearings-only SLAM," in *Proc. IEEE Int. Conf. Robot. Automat.*, Barcelona, Apr. 2005, pp. 2422–2427.

[14] M. Bryson and S. Sukkarieh, "An information-theoretic approach to autonomous navigation and guidance of an uninhabited aerial vehicle in unknown environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Edmonton, Aug. 2005.

[15] A. Adam, E. Rivlin, and I. Shimshoni, "Computing the sensory uncertainty field of a vision-based localization sensor," *IEEE Trans. Robot. Automat.*, vol. 17, no. 3, pp. 258–267, Jun. 2001.

[16] T. Broida, S. Chandrashekhar, and R. Chellappa, "Recursive 3-d motion estimation from a monocular image sequence," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 26, no. 4, pp. 639–656, Jul. 1990.

[17] T. Bailey, "Constrained initialisation for bearing-only SLAM," in *Proc. IEEE Int. Conf. Robot. Automat.*, vol. 2, Taipei, Sep. 2003, pp. 1966–1971.

[18] D. J. C. MacKay, "Information based objective functions for active data selection," *Neural Comput.*, vol. 4, no. 4, pp. 589–603, 1992.

[19] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*. New York: John Wiley & Sons, 2001.

[20] R. Sim and N. Roy, "Global A-optimal robot exploration in SLAM," in *Proc. IEEE Int. Conf. Robot. Automat.*, Barcelona, Apr. 2005, pp. 673–678.

[21] J. Andrade-Cetto and A. Sanfeliu, "The effects of partial observability when building fully correlated maps," *IEEE Trans. Robot.*, vol. 21, no. 4, pp. 771–777, Aug. 2005.