Incremental Structure From Restrictive Motion

Abstract

This paper demonstrates the ability of the Federated Information Sharing filter (FIS) to causally extract 3D structure and camera motion from monocular image sequences when the camera motion is singular, *ie.* restricted to straight trajectories heading to the objects to be modelled. A special Multi-Hypothesis, Bearings-Only, Simultaneous Localisation And Mapping scheme based on Extended Kalman Filter is used: its capability of immediate initialisation of multiple simultaneous landmarks, based on FIS, is the key to cope with the weak observability of the system. All the image processing to extract the pertinent visual information in real time is described. Experimental results show the performance of the proposed methods.

1 Introduction

In this paper we address the real time, incremental solution to the Structure From Motion (SFM) problem with singular motions. A camera being a bearings-only sensor, depth information of the perceived features is not observable from a single picture shot. If we add camera motion, this depth observability becomes strongly dependent on the performed trajectory. In cases where this trajectory cannot be arbitrarily chosen, the question of whether SFM can still be solved or not should normally be given a negative reply.

The most pathological case (Fig. 1) results from moving along a straight trajectory in the same direction the camera is looking at. In the autonomous vehicles field, this *singular motion* situation turns out to be very common: every vehicle has interest in looking at the place it is heading to. Depth recovery and hence 3D reconstruction of the perceived scene in such a situation (which would be of great interest for navigation tasks) is all but a trivial issue.

Numerous authors build off-line a model, as a set of 3D points, from an image sequence, using the Bundle Adjustment algorithm [8]: then the model is exploited in robotics applications to cope with the robot self-localisation [10, 4]. Other authors [3, 1, 6] have already addressed the problem of incrementally (or causally) extract 3D structure plus motion from monocular vision in real time. They have normally used a Simultaneous Localisation And Mapping (SLAM) approach (see [12] for the seminal paper, and numerous contributions using Particle Filters, Information Filters, etc.) using an Extended Kalman Filter (EKF) as the fusion engine, that naturally provides incremental operation. In these Bearings-Only SLAM approaches, observations are interest points [5, 9, 4] extracted from images acquired by the moving camera; the world model is represented as a sparse set of 3D points called *landmarks*. Observations are causally used to: *a)* include or *initialise* new landmarks; *b)* refine or *correct* this world model; and *c)* get *localised* in it, hence recovering the camera motion. We'd like to highlight the work by Davison [3] as a cornerstone in the subject. Chiuso *et al.* [2] made a well theoretically founded work

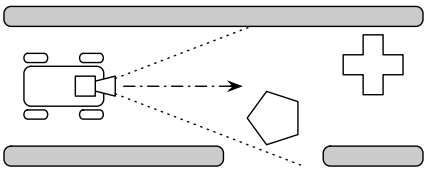

Figure 1: Quasi singular motion for SFM

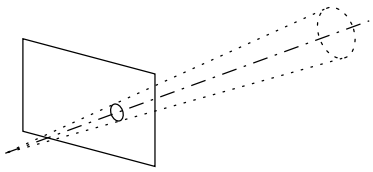

Figure 2: The conic ray extends to infinity

where they present a very interesting way to minimise the (inevitable) drift of the (unobservable) scale factor. However, when put into work in such a pathological case, none of these algorithms will succeed in recovering depths for landmarks that are close to the motion axis: the point of view evolves too slowly to permit observability in the limited duration of the initialisation process they use.

More recently, Solà *et al.* [13] and Lemaire *et al.* [7] have proposed new, more optimised and complementary ways to improve the landmarks initialisation procedure, which is one of the main difficulties when performing SLAM with monocular vision. When a new landmark is detected in the image, the back-projection of the noisy measured pixel defines a conic *pdf* for the landmark position, called *ray*, that extends to infinity (Fig. 2). This ray is first truncated at minimum and maximum considered depths, and then approximated by a geometric series of Gaussians. The terms of this series are successively pruned as new observations make their likelihoods drop below a certain threshold. The landmark's depth is considered observed when only one term is left. These approaches are complementary in the sense that, in [7], landmark initialisation into the SLAM map takes place at the end of this pruning process, while in [13] all terms are initialised from the beginning.

This latter undelayed initialisation method [13] has two advantages that can be exploited to unblock our problem: *a)* it allows an indefinite amount of time to initialise a landmark; and *b)* during this time, and because the landmark is already mapped, it can exploit the bearing information that these features do provide to improve camera localisation. Further, the proposed Federated Information Sharing (FIS) technique allows us to initialise several landmarks simultaneously.

Perception issues related to feature tracking or matching arise when working under such singular trajectories. Visual features, typically interest points extracted by the Harris detector [5], suffer from important changes in appearance, something that did not happen in [3], where matching could be achieved with simple correlations with a fixed patch. For better results Shi and Tomasi [11] or Lowe's SIFT features [9] are good alternatives, but they demand considerably higher computational load. We chose to add some light improvements to the matching method in [3] to obtain a more robust operation while keeping the same low computational cost.

This paper proceeds as follows. In section 2 we review the FIS technique for undelayed landmarks initialisation. In section 3 we present the proposed methods for features selection and matching. Experimental results are exposed in section 4 and conclusions come in section 5.

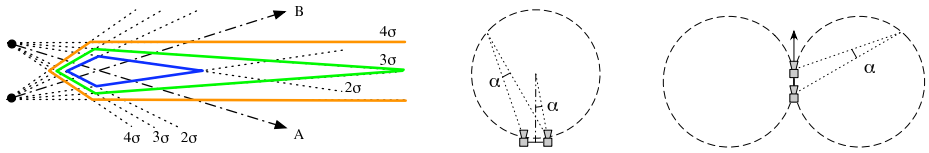

Figure 3: Different regions of intersection for 4σ (orange), 3σ (green) and 2σ (blue) ray widths (*left*). The angle between rays axes A and B is $\alpha = 8\sigma$, which is used to draw the simplified depth observability regions in a stereo head (*center*) and a camera travelling forward (*right*).

2 Bearings-Only SLAM with singular trajectories

2.1 Observability issues

Just a couple of ideas (not to be strictly interpreted) to help to understand the observability concept that we use (Figs. 2 and 3). Let us consider two features extracted from two images and matched because they correspond to the same landmark: their back-projections are two conic rays that extend to infinity. These ray's angular widths can be defined as a multiple of the standard deviation σ of the angular error, which depends on the camera angular resolution (pixel size, focal length), on the camera pose uncertainty, and on the accuracy of the point detector. We say that the landmark's depth is observed if the region of intersection of these rays is *a)* closed and *b)* sufficiently small. If we impose, for example, the two external 4σ bounds of the rays to be parallel, then we assure that the 3σ intersection region (which covers 99% probability) is closed and that the 2σ one (covering 95%) is small. The angle α between the two rays axes is then $\alpha = 8\sigma = constant$.

We can plot the geometric place of those world points for which the two angular observations differ exactly in this angle α. Inside the obtained circle, depth is observable; outside it is not. Given an overall angular uncertainty σ, this region's radius is directly proportional to the distance between the two cameras. In 3D, the region is radially located around the axis joining both cameras, producing something like a toroidal region. In a stereo configuration or for a lateral motion of a moving camera like in aerial images (Fig. 3 *center*), this region is in front of the sensor. This paper considers only singular motions, i.e. a single camera moving forward (Fig. 3 *right*); precisely in the singular direction depth recovery is simply impossible. Close to and around this axis, observability is only possible if the region's radius becomes very large. This implies the necessity of very large displacements of the camera during the initialisation process, something that can only be accomplished with undelayed initialisations.

2.2 Federated Information Sharing SLAM

The core of the algorithm is an EKF-based SLAM system, which is not explained here. We review here the FIS undelayed initialisation technique proposed in [13] and refer the reader to this reference for more precise explanations. A non negligible contribution of this paper is the addition of a *re-balance* step on the likelihoods that will guarantee correct ray's pruning in very long initialisations.

2.2.1 Rays initialisation

At the first observation of a landmark, its *pdf* (the conic ray) is approximated with a geometric series of Gaussians. Each Gaussian i represents an hypothesis for the position of the landmark, and is assigned a depth mean $\hat{s}_i$ and standard deviation σ_i using the geometric series rule:

$$\begin{aligned} \sigma_i &= \alpha \cdot \hat{s}_i \\ \hat{s}_{i+1} &= \beta \cdot \hat{s}_i \end{aligned} \tag{1}$$

where α is the aspect ratio and β the geometric base. Each one of these hypotheses is initialised in the SLAM map with the standard EKF-SLAM method, considering the tuple $\{\hat{s}_i, \sigma_i^2\}$ as the depth observation's mean and variance. Ray's hypotheses initial likelihoods Λ_i are uniformly distributed:

$$\Lambda_i = 1/N_g \quad ; \quad 1 \leq i \leq N_g \tag{2}$$

where N_g is the initial number of hypotheses. This number is between 3 and 7 [13], depending on our knowledge on the minimal and maximal depths for the landmarks.

2.2.2 Rays FIS updates

Subsequent observations of these landmarks are used to *1)* update likelihoods for each hypothesis; *2)* prune unlikely hypotheses; and *3)* correct the map and the camera localisation. A further operation of *4)* re-balancing likelihoods is added to improve ray stability in very long initialisations.

1. Likelihood update. Member's likelihoods λ_i are computed with respect to the current observation as in [13]. The aggregated likelihood Λ_i is the product of all likelihoods of a member since it was initialised: $\Lambda_i(t) = \Pi_{\tau=0}^{t}\big(\lambda_i(\tau)\big)$. Hence it is updated with the equivalent incremental formula $\Lambda_i(t) = \lambda_i(t)\Lambda_i(t-1)$ that we will write:

   $$\Lambda_i^+ = \Lambda_i \cdot \lambda_i \tag{3}$$

   where the notation Λ^+ stands for *the updated value of* Λ. All aggregated likelihoods are then normalised to $\Sigma(\Lambda_i) = 1$.

2. Members pruning. Those hypotheses with aggregated likelihood smaller than a certain threshold are pruned:

   $$\Lambda_i < \tau/N \Rightarrow \textit{prune member } i \tag{4}$$

   where τ is a fixed threshold and N is the ray's current number of members. If one single member remains after pruning, the ray switches to the category of *point* with no other special thing to consider. Points are corrected as in the standard EKF-SLAM algorithm.

3. FIS update. Map correction with surviving members is processed with the FIS method to avoid inconsistency. The observation's information (the inverse of the covariances matrix $\mathbf{R}$) is shared among all the hypotheses proportionally to their aggregated likelihood:

   $$\mathbf{R}_i^{-1} = \Lambda_i \cdot \mathbf{R}^{-1} \Rightarrow \mathbf{R}_i = \mathbf{R}/\Lambda_i. \tag{5}$$

An EKF-SLAM update is then performed over each hypothesis using the same measure but with the corresponding modified covariances matrix $\mathbf{R}_i$.

4. Likelihoods re-balance. In multi-hypotheses filters in the presence of Gaussian noises, the distribution of weights Λ_i of the set of hypotheses can be proven to degenerate as time tends to infinity. This means that one hypothesis will take the whole weight while the others will consequently vanish. This is in fact an advantage for observable systems as the filter will probably tend to a single-hypothesis one (*i.e.* an EKF!), the choice of the right hypothesis being driven by the observations. But it is catastrophic for weakly observable systems, specially if one wishes to prune the weakest hypotheses as we do. If the depth of the landmark is not observable for a certain amount of time, we are easily going to prune the right hypothesis, something that can not be undone. This degeneracy must definitely be avoided. It can be done by vanishing the effect of very old observations on the aggregated likelihood so as to have $\Lambda_i(t) = \Pi_{\tau=0}^{t}\big(\lambda_i(\tau)\big)^{(1-\gamma)(t-\tau)}$ instead of the uniform $\Lambda_i(t) = \Pi_{\tau=0}^{t}\big(\lambda_i(\tau)\big)$ foreseen. Most of the work being done incrementally by (3), we just need to add this re-balance operation:

$$\Lambda_i^+ = \Lambda_i^{(1-\gamma)}, \tag{6}$$

where $0 \leq \gamma \leq 1$ is called the *vanishing factor*.

This FIS method, allowing undelayed initialisation of landmarks in the stochastic map, has been validated by simulations presented in [13], considering mainly loops, to deal with the classical *loop closing* problem in SLAM. Here this method is validated with actual sensory data acquired from a camera mounted on a robot going forward. Before showing experimental results, perception issues (feature detection and matching) are considered in the next section.

3 Perception

3.1 Landmarks model

Each perceived landmark is modelled by both geometrical and photometrical information. The geometrical part consists in its Euclidean position in space, $\mathbf{x}_p = \begin{bmatrix} x_p & y_p & z_p \end{bmatrix}^\top$, which is stochastically included in the SLAM map with the mentioned FIS initialisation method.

The photometrical part is an appearance-based model consisting of a medium-sized patch around the image of the point. This model, used in many real time vision-based SLAM works, is suitable for posterior feature matching by a maximum of correlation scan inside the predicted ellipses (sometimes referred as *active search*).

However, because of the performed trajectory, our application induces considerable changes in appearance (orientation, deformation and scale) and soon the registered patch will poorly describe the current appearance of the landmark. Because a durable observation of a landmark is a must for succeeding in its initialisation, we must guarantee a more robust model. Our solution is as simple as re-registering a new patch before its degradation becomes too important. To avoid patch refreshing during occlusions, we also

require that this degradation be slow. This can be easily accomplished with the following *degradation test*: If the mean of the last n scores is between LO and HI (which means the patch is moderately getting degraded), and its standard deviation is lower than TH (which means it does it at a slow rate), then we register a new patch with the current landmark's image. We use the Zero-mean, Normalised Correlation Score (ZNCC) for correlations. For the values of n, LO, HI and TH refer to the Experiments in section 4.

3.2 Feature detection for initialisation

New landmarks to be initialised correspond to the strongest Harris points [5] in the less populated image regions. Some heuristic (but not relevant) reasoning must be made to limit *a)* the size of these regions; *b)* their number and distribution; *c)* the number of landmarks per region; and *d)* the total number of simultaneous initialisations. All of them will depend on the particularities of the application and on current computation capacities. We use (Fig. 4) a fixed grid that divides the image in a number of equal regions, and impose that a minimum number of landmarks (not necessarily the same) should exist inside each one of them.

3.3 Selection of existing points and rays to measure

Subsequent observations are performed only on a selected set of points and rays, a *point* being a landmark that is represented by a single hypothesis (hence considered totally observed). This selection is an important step to satisfy drastic real time constraints for a robotics application: landmarks will be used to estimate the camera motion. First, uncertainties of all landmarks are projected onto the image plane (Fig. 4). This projections result in one ellipse per point (normally the 2σ or 3σ bound ellipses are taken) and a set of overlapping ellipses per ray. The determinant of each ellipse is taken as a comparative scalar measure of its surface. The larger the surface, the bigger the information gain when performing an observation to that landmark, the greater the interest in measuring it. For rays, and because ellipses overlapping and likelihoods weighting complicate this information gain measure, we will take the biggest of all the determinants as a simplified measure: as long as we compare points against points and rays against rays, this simplification will do.

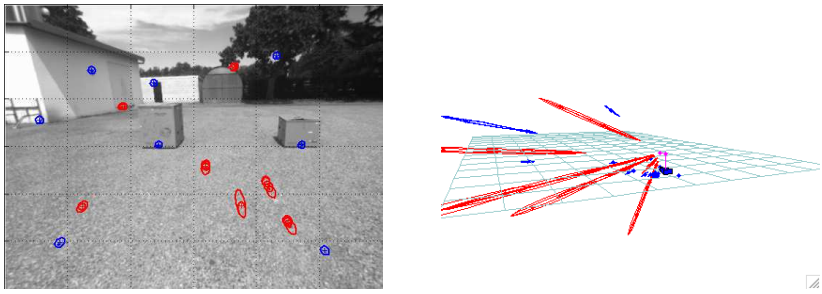

Figure 4: Image regions for new features search (*dotted grid*). Predicted ellipses for points (*blue*). Predicted overlapping ellipses for rays (*red*). Successful measures are plotted as tiny dots inside the predicted regions. The corresponding 3D map is shown.

The update sequence proceeds in two main steps: *1)* A limited number n_p of points corresponding to the largest ellipses are corrected using the EKF-SLAM correction step. *2)* A limited number n_r of rays corresponding to the largest ellipses are corrected using the FIS-SLAM update step (which includes likelihood update, pruning and correction). Proceeding in this order provides *a)* the best possible camera relocation with good 3D information (the points), that afterwards *b)* allows the ray's members to be more easily discriminated and pruned, while *c)* minimising the effect of the FIS correction on wrong hypotheses, hence *d)* preventing inconsistency and divergence problems to occur.

3.4 Performing the landmark measure

The measure is defined as the u and v co-ordinates of the point in the image whose associated patch best correlates with the registered patch of the landmark. This is accomplished in four steps (Fig. 5): *1)* The strongest Harris points are extracted inside the predicted ellipse (or set of ellipses) region; *2)* The one that best correlates with the landmark patch is retained; *3)* A final local search in the close neighbourhood of this pixel is performed to get the maximum correlation score; and *4)* A sub-pixellic result is obtained by computing the scores of the four cardinal neighbours of the best pixel, and locating the maximums of the parabolic interpolations in the horizontal direction (obtaining the u sub-pixellic co-ordinate) and the vertical one (for the v co-ordinate).

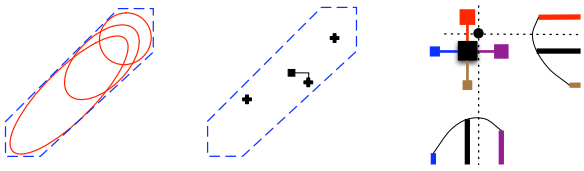

Figure 5: Measure extraction. Ellipses or sets of ellipses define a region. Harris points are extracted inside this region. A local search is performed around the best one. Parabolic interpolations with the four cardinal neighbours gives a sub-pixellic measurement.

4 Experiments

Different experiments in real scenarios have been carried out. A representative one corresponds to that in Fig. 4, in which a vehicle (an all terrain rover) performs a straight trajectory that passes between the two boxes. The motion axis corresponds roughly to the central point of the horizon line in the image, slightly above the lower left corner of the round hut at the further end. Observability is very weak in the four image regions surrounding this point. It gets better as we move away from it. The bottom regions are perfectly observable, but the rapid changes in appearance difficult matching there, lowering feature stability.

The different parameters are tuned as follows. *1)* For the rays initialisation and updates scheme (section 2.2) we take $\{\alpha,\beta,\tau,\gamma\} = \{0.3,3,0.01,0.2\}$, with $\hat{s}_1 = 1$ and $N_g = 4$, which gives a ray range from 1m to well beyond 30m. *2)* For the photometrical landmark model (section 3.1) we take patches of 15x15 pixels and tune the model degradation test

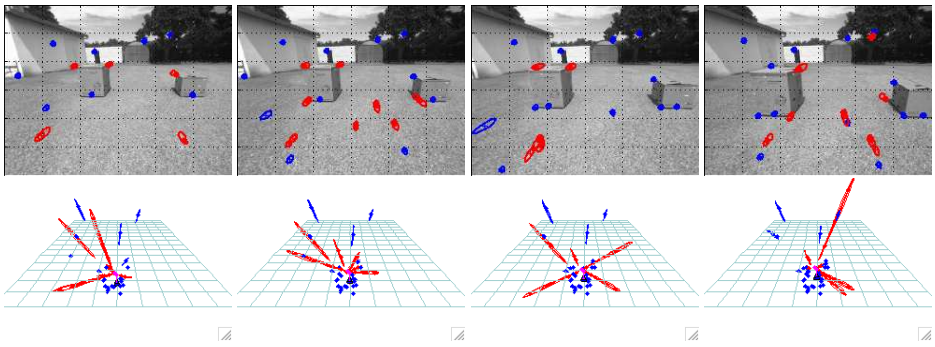

Figure 6: A portion of the reconstruction sequence at 5 frames (approx. 35cm) intervals. Image plane and 3D reconstruction are shown for frames 46, 51, 56 and 61.

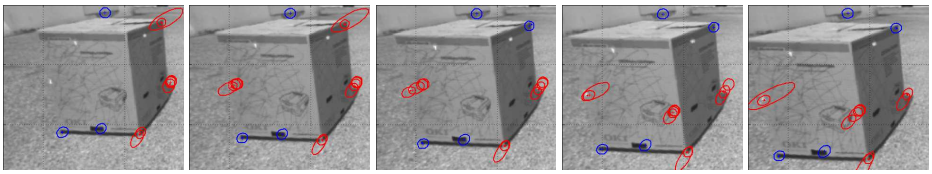

Figure 7: A zoom of another portion of the reconstruction sequence at 1 frame (approx. 7cm) intervals. Image plane is shown for frames 64 to 68.

with the values $\{n, LO, HI, TH\} = \{4, 0.8, 0.9, 0.1\}$. *3)* For the image regions for new initialisations (section 3.2), we take a grid of 6*x*6 regions and impose at least one landmark at each one of the 4*x*4 inner regions. A maximum of 8 new initialisations per frame is allowed, and the overall number of running initialisations is limited to 12. *4)* For the selection of existing points and rays to correct (section 3.3) we take $\{n_p, n_r\} = \{8, 10\}$. *5)* To perform the measures (section 3.4), we take the 2σ-bound projected ellipses, that enclose a region of 95% probability of actually containing the searched feature.

For this 3D experiment we disposed of the robot's 2D odometry, which we used to improve predictions and, more important, to fix the scale factor. We chose for it a very simple error model in which translation noise variances σ_x^2, σ_y^2 and σ_z^2 and rotation (the three Euler angles) noise variances σ_ϕ^2, σ_θ^2 and σ_ψ^2 are proportional to the performed forward displacement Δx:

$$\begin{aligned} \sigma_x^2 = \sigma_y^2 = \sigma_z^2 &= {k_d}^2 \cdot \Delta x \\ \sigma_\phi^2 = \sigma_\theta^2 = \sigma_\psi^2 &= {k_a}^2 \cdot \Delta x \end{aligned} \tag{7}$$

with $k_d = 0.04\text{m}/\sqrt{\text{m}}$ and $k_a = 0.02\text{rad}/\sqrt{\text{m}}$.

The sequence consists of 97 images, taken at $\Delta x = 7$cm intervals approximately. Fig. 6 shows a snapshots sequence at 5 frames intervals. Observe how rays are initialised and their convergence to single points at the lower corners of both boxes (with fair observability) while at the upper ones (weak observability) the initialised rays remain as rays. They are used to improve camera (and robot!) localisation thanks to the undelayed initialisation.

The rays initialisation and pruning based on reasoning in the image plane can be better

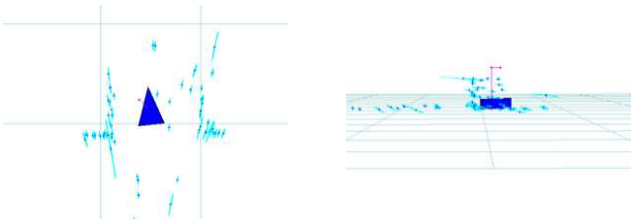

Figure 8: Top and side views of the reconstructed boxes at the end of the sequence.

appreciated in the zoomed sequence of consecutive snapshots shown in Fig. 7.

A systematic error of about $0.6^\circ/\text{m}$ was perceived in the angular odometry readings: while the robot was slightly turning left, the integration of the odometry data results in a slow turn to the right. The 3D reconstruction did not suffer from this angular bias and the final estimated trajectory represents well the real one, *i.e.* it turns to the left (note that precise ground truth is not available).

The reconstructed structure of the two boxes is shown in Fig. 8. On the top view the structure of the boxes is clearly visible. On the side view, points on the ground are also well reconstructed. The recovery of the scale factor is shown to be correct as these ground points are effectively on the ground: they are consistent with the fact that the camera is at a precise height of 1.02m.

5 Conclusions and extensions

This paper has presented a Bearing-Only SLAM method, i.e. an incremental way to tackle the Bundle Adjustment algorithm to solve the Structure From Motion problem. Many improvements can be made to the matching scheme (section 3.1), but careful and accelerated realisations will have to be implemented if more robust features [11, 9] have to be used, so that a higher feature density could be extracted from poorly textured images. The presented technique has been validated on numerous image sequences: the prediction-scan-match procedure inside the predicted ellipse (or set of ellipses for rays) allows a rapid rejection of wrong matches, and the ability to simultaneously initialise several landmarks compensates for any eventual loss of features during this phase.

The FIS strategy allows an undelayed initialisation for new detected landmarks, so that our method is efficient even for singular motions of the camera. This is specifically required when the camera is mounted on a vehicle moving along a straight line (highway or urban navigation). Experimental results have shown that several hypothesis for a single ray can remain active in the map during numerous iterations, mainly for rays almost aligned with the camera trajectory, created by landmarks far from the sensor.

By now two extensions of this method are being implemented:

1) A second camera is added to the system. For points that are observable based on stereo-vision triangulation, the initialisation is immediate. Farther away, landmarks are initialised as rays, but starting at the limit of the stereo observable distances. If this distance reaches, say, 50m, the ray's first term is initialised there; with $\beta = 3$ the other terms are at 150m, 450m and 1350m. We get one kilometre with just 4 hypotheses. Further, if this landmark gets eventually occluded in one camera, it can be still measured

with the other one, updating the system accordingly.

2) In the presence of moving objects in the scene, the multi hypotheses capabilities of the FIS filter will be exploited to estimate their trajectories, which need as-well very long observations to be determined. This will lead to visual-based SLAMMOT, standing for SLAM with Moving Objects Tracking.

References

[1] T. Bailey. Constrained initialisation for bearing-only slam. *IEEE International Conference on Robotics and Automation*, 2003.

[2] A. Chiuso, P. Favaro, H. Jin, and S. Soato. Structure from motion casually integrated over time. In *IEEE transactions on pattern analysis and machine intelligence*, 2002.

[3] A.J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proc. International Conference on Computer Vision, Nice*, October 2003.

[4] L. Goncalves and al. A visual front-end for simultaneous localization and mapping. In *Int. Conf. on Robotics and Automation (ICRA'05)*, 2005.

[5] C. Harris and M. Stephens. A combined corner and edge detector. In *Fourth Alvey Vision Conference, Manchester (UK)*, 1988.

[6] N. M. Kwok and G. Dissanayake. An efficient multiple hypothesis filter for bearing-only slam. In *IEEE/SRJ International Conference on Intelligent Robots and Systems*, Sendai, Japan, 2004.

[7] Thomas Lemaire, Simon Lacroix, and Joan Solà. A practical 3d bearing only slam algorithm. In *IEEE International Conference on Intelligent Robots and Systems*, august 2005.

[8] M.I.A. Lourakis and A.A. Argyros. The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm. Technical Report 340, Institute of Computer Science - FORTH, Heraklion, Crete, Greece, 2004.

[9] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60, 2*, pages 91–110, 2004.

[10] S. Se, D.G. Lowe, and J. Little. Global localization using distinctive visual features. In *Proc. Int. Conf. on Intelligent Robots and Systems (IROS'02)*, pages 226–231, 2002.

[11] J. Shi and C. Tomasi. Good features to track. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR'94), Seattle, USA*, pages 593–600, 1994.

[12] R. Smith and P. Cheeseman. On the representation and estimation of spatial uncertainty. *The International Journal of Robotics Research*, 5(4):56–68, 1987.

[13] Joan Solà, André Monin, Michel Devy, and Thomas Lemaire. Undelayed initialization in bearing only slam. In *IEEE International Conference on Intelligent Robots and Systems*, august 2005.