2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Inertial Aiding of Inverse Depth SLAM using a Monocular Camera

Pedro Piniés, Todd Lupton, Salah Sukkarieh, Juan D. Tardós

Abstract— **This paper presents the benefits of using a low cost inertial measurement unit to aid in an implementation of inverse depth initialized SLAM using a hand-held monocular camera. Results are presented with and without inertial observations for different assumed initial ranges to features on the same dataset. When using only the camera, the scale of the scene is not observable. As expected, the scale of the map depends on the prior used to initialize the depth of the features and may drift when exploring new terrain, precluding loop closure. The results show that the inertial observations help to improve the estimated trajectory of the camera leading to a better estimation of the map scale and a more accurate localization of features.**

I. Introduction

Performing Simultaneous Localization and Mapping (SLAM) using a monocular camera has received a lot of attention over the past few years [1][2][3][4]. Many successful implementations have been demonstrated which work in limited situations. The main limitation of SLAM with bearing only sensors is that the scale factor of the map is not observable as there is no measurement of the range from the camera to the landmarks. This is addressed in some systems by initializing the system looking at a pattern of known size [1]. However, as this is the only measurement of distance, the scale of the map may drift when exploring new areas, making loop closure difficult.

In most bearing-only SLAM systems, the initialization of new features in the map is delayed until there is enough parallax to estimate the depth of the features. The work presented in [4] highlights the differences between delayed and undelayed initializations, the benefits of undelayed methods and its implementation difficulties. A recent paper [5] has proposed a new technique for undelayed initialization of features using the inverse depth of the features relative to the camera position from where the feature was first observed. This technique is able to deal in an uniform way with close and far features from the first instant they are detected. In particular, far features provide very useful information about the camera orientation, reducing the angular drift in long motions. These characteristics make it desirable for use in large unstructured environments. However, the undelayed initialization technique uses a prior to initialize the inverse depth of the landmarks (in the original paper, a Gaussian distribution with mean $0.5m^{-1}$ and standard deviation $0.5m^{-1}$). As this initial prior is the only information of distance used by the system, it introduces a bias in the size of the final map obtained. This lead not only to ambiguity in the overall scale factor of the map, which is inevitable in a monocular vision system, but also to varying scale factors in the relative range between features in the map, which can make loop closure impossible. The initial range estimate is also used as a linearization point for subsequent observations, which can introduce linearization errors leading to filter inconsistency.

In this paper we demonstrate the benefit of using a low cost inertial measurement unit (IMU) to aid in an inverse depth implementation of bearing only SLAM. It is shown that the inertial observations constrain the uncertainty of the camera location leading to more accurate initialization of features. Furthermore there will be less variation in the scale factor between features in the map which produces a more consistent map with less uncertainty. This is akin to the property of geometric similarity, congruency to a scale factor. Another benefit of inertial measurements, not explored in this paper, is its ability to provide more accurate predictions of the location of features in the next image, improving the robustness and efficiency of data association.

Section II of this paper gives a brief overview of inertial SLAM as well as inverse depth initialization and the observation model used. Section III describes the experimental setup and the characteristics of the sensors. Section IV presents and discusses the results obtained and Section V provides a conclusion and indication of future work.

II. Inertial SLAM

A. Brief IMU description

An inertial measurement unit (IMU) can be a valuable sensor in many applications since it provides information about the movement of the vehicle it is attached to independently of the characteristics of the platform. The IMU provides measurements of its own acceleration and angular velocity at high update rates. From them position, velocity and attitude of the platform can be calculated via integration.

However there are disadvantages when using an IMU, especially for low-cost units as the ones described in this paper. The errors in the estimation are fundamentally caused by the bias in the accelerometers and gyros and the random walk produced by integration of the intrinsic noise in them [6]. Gyros errors have the most detrimental effect since the attitude calculated is used to compute and cancel the gravitational acceleration on the observed accelerations. If the platform accelerations are smaller than the gravitational ones then even small errors in attitude produce significant drifts in the velocity and position estimates. To compensate

Pedro Piniés and Juan D. Tardós are with Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, María de Luna 1, E-50018, Zaragoza, Spain,{ppinies, tardos}@unizar.es

Todd Lupton and Salah Sukkarieh are with ARC Centre for Excellence in Autonomous Systems, Australian Centre for Field Robotics, University of Sydney, {t.lupton, salah}@cas.edu.au

these errors external information gathered by other sensors is required. Further details can be found in [7].

B. State vector

In inertial SLAM [8] the vehicle position, velocity and attitude and a map with the most relevant feature locations of the environment are estimated using relative information between the vehicle and each feature. The state vector to be estimated is then given by:

$$\mathbf{x}(k) = \begin{bmatrix} \mathbf{x}_v(k) \\ \mathbf{Y}(k) \end{bmatrix} \tag{1}$$

where $\mathbf{x}_v(k)$ represents the vehicle state

$$\mathbf{x}_v(k) = \begin{bmatrix} \mathbf{r}^n(k) \\ \mathbf{v}^n(k) \\ \mathbf{\Psi}^n(k) \\ \mathbf{f}^b_{bias}(k) \\ \omega^b_{bias}(k) \end{bmatrix} \tag{2}$$

and $\mathbf{Y}(k)$ the set of n features in the map

$$\mathbf{Y}(k) = \begin{bmatrix} \mathbf{y}_1(k) \\ \vdots \\ \mathbf{y}_n(k) \end{bmatrix} \tag{3}$$

The components of the features $\mathbf{y}_i(k)$ will be described in II-D. The vehicle state $\mathbf{x}_v(k)$ contains the three cartesian coordinates of the vehicle position $\mathbf{r}^n$, velocity $\mathbf{v}^n$ and attitude in Euler angles $\mathbf{\Psi}^n$, all of them represented with respect to the navigation frame N and the bias in the accelerometers $\mathbf{f}^b_{bias}$ and gyros ω^b_{bias} in the body frame B.

C. Process model

The dynamic evolution of the state in time is given by a non-linear state transition function:

$$\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k), \mathbf{u}(k), \mathbf{w}(k)) \tag{4}$$

where the input $\mathbf{u}(k)$ encloses the body-frame referenced accelerations $\mathbf{f}^b(k)$ and angular velocities $\omega^b(k)$ measured by the IMU

$$\mathbf{u}(k) = \begin{bmatrix} \mathbf{f}^b(k) \\ \omega^b(k) \end{bmatrix} \tag{5}$$

and the term $\mathbf{w}(k)$ represents the noise in those measurements as a zero mean uncorrelated gaussian noise with covariance $\mathbf{Q}$

$$\mathbf{w}(k) = \begin{bmatrix} \delta\mathbf{f}^b(k) \\ \delta\omega^b(k) \end{bmatrix} \tag{6}$$

The evolution of the vehicle state given the previous input and noise can be calculated using the following equations:

$$\begin{bmatrix} \mathbf{r}^n(k+1) \\ \mathbf{v}^n(k+1) \\ \mathbf{\Psi}^n(k+1) \\ \mathbf{f}^b_{bias}(k+1) \\ \omega^b_{bias}(k+1) \end{bmatrix} =$$

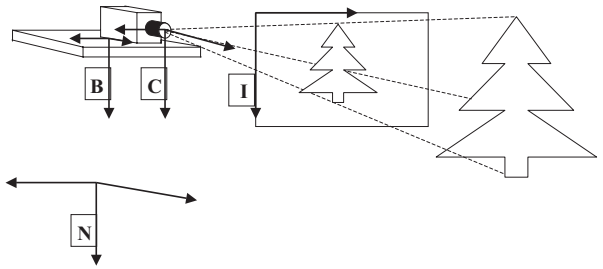


Fig. 1. Frames description.

$$\begin{bmatrix} \mathbf{r}^n(k) + \mathbf{v}^n(k+1)\Delta t \\ \mathbf{v}^n(k) + \left[C^n_b(k)[(\mathbf{f}^b(k) + \delta\mathbf{f}^b(k)) - \mathbf{f}^b_{bias}(k)] + \mathbf{g}^n\right]\Delta t \\ \mathbf{\Psi}^n(k) + E^n_b[(\omega^b(k) + \delta\omega^b(k)) - \omega^b_{bias}(k)]\Delta t \\ \mathbf{f}^b_{bias}(k) \\ \omega^b_{bias}(k) \end{bmatrix}$$

where $C^n_b(k)$ and E^n_b are the direction cosine matrix and rotation rate transformation matrix respectively [8]. The biases are assumed to be constant and affected by gaussian noise.

It can be observed in the velocity prediction that the gravitational acceleration is canceled out on the observed accelerations $\mathbf{f}^b(k)$ by adding the term $\mathbf{g}^n$ which is the gravity vector in the navigation frame. As stated previously, if the error in the attitude estimate (C^n_b) is incorrect the compensation achieved by adding $\mathbf{g}^n$ will be incorrect and the velocity and position estimates will drift.

D. Measurement equation

The method implemented to deal with the undelayed bearing only SLAM is an adaptation of the inverse depth algorithm proposed in a recent paper [5]. According to this method the representation of the features is over-parametrized as follows:

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{r}^n_i \\ \theta^n_i \\ \phi^n_i \\ \rho^n_i \end{bmatrix} \tag{7}$$

where $\mathbf{r}^n_i$ represents the camera optical center, in cartesian coordinates, from where the feature was first observed. The angles θ^n_i, ϕ^n_i define the azimuth and elevation of the ray that goes from the initial camera position to the 3D point feature. Finally $\rho^n_i = 1/d_i$ is the inverse of the distance d_i between that camera position and the feature.

The frames involved in the prediction equation can be seen in Fig. 1, where the navigation frame is represented by N, the body frame associated to the inertial by B, C is the frame of the camera which is stuck on the inertial and I is the reference frame associated with the image plane.

In our implementation of the inverse depth the observa-

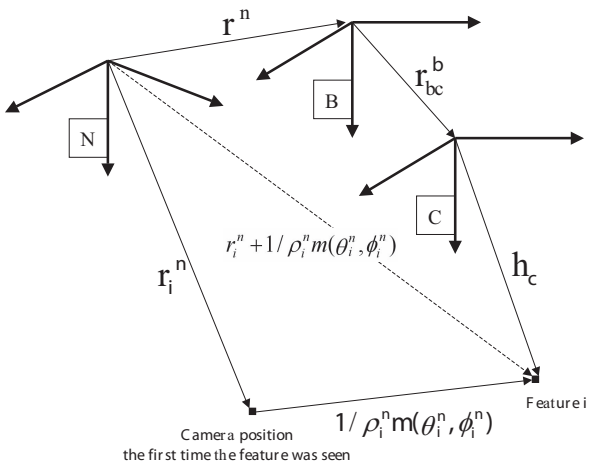


Fig. 2. Graphical representation of Eq(9). It shows the relations between the components of the vehicle state and the feature i to obtain the ray vector $\mathbf{h}$ that goes from the camera center to the 3D position of feature i.

tions are related to the state by:

$$\mathbf{z}_i = \begin{bmatrix} \theta_i^c \\ \phi_i^c \end{bmatrix} = \mathbf{h}(\mathbf{r}^n, \mathbf{\Psi}^n, \mathbf{r}_i^n, \theta_i^n, \phi_i^n, \rho_i^n) + \mathbf{v} = \begin{bmatrix} \arctan\left(\frac{h_y^c}{h_x^c}\right) \\ \arctan\left(\frac{h_z^c}{\sqrt{(h_x^c)^2+(h_y^c)^2}}\right) \end{bmatrix} + \mathbf{v} \quad (8)$$

where $[h_x^c, h_y^c, h_z^c]^T$ are the components of the vector $\mathbf{h}^c$ which defines the ray that goes from the current camera position to the 3D point in camera coordinates (^{c}) and $\mathbf{v}$ is the uncorrelated, zero-mean gaussian observation noise with covariance $\mathbf{R}$.

Equation (9) shows how to calculate the ray $\mathbf{h}^c$ from the components of the vehicle state $\mathbf{x}_v$ and the corresponding feature $\mathbf{y}_i$. The term $\mathbf{r}_{bc}^b$ is the sensor offset from the inertial measured in the body frame and the matrix C_b^c is the transformation matrix from the body frame to the camera frame. Fig.2 offers a graphical representation of this equation.

$$\begin{aligned} \mathbf{h}^c &= C_b^c C_n^b \mathbf{h}^n \\ \mathbf{h}^n &= \left(\mathbf{r}_i^n + \frac{1}{\rho_i^n}\mathbf{m}(\theta_i^n, \phi_i^n)\right) - (\mathbf{r}^n + C_b^n \mathbf{r}_{bc}^b) \end{aligned} \quad (9)$$

The vector $\mathbf{m}$ in Eq(9) is a unitary vector that describes the direction of the ray when the feature was seen for the first time. It can be calculated from the azimuth θ_i^n and elevation ϕ_i^n angles of the feature by:

$$\mathbf{m}(\theta_i^n, \phi_i^n)) = \begin{bmatrix} \cos(\theta_i^n)\cos(\phi_i^n) \\ \sin(\theta_i^n)\cos(\phi_i^n) \\ \sin(\phi_i^n) \end{bmatrix} \quad (10)$$

After applying an undistortion process to the points of interest in the image a pinhole camera model is used to determine the azimuth and elevation angles in the camera frame from the pixel coordinates (u, v) of the feature.

$$\begin{bmatrix} \theta_i^c \\ \phi_i^c \end{bmatrix} = \begin{bmatrix} \arctan\left(\frac{u-u_0}{f_u}\right) \\ \arctan\left(\frac{v-v_0}{f_v}\right) \end{bmatrix} \quad (11)$$

where u_0,v_0 are the center coordinates in the image and f_u, f_v are the components of the focal length.

Fig. 3. Picture of the inertial and camera used in the experiments.

III. Experimental Setup

The situation in which bearing only initialisation techniques have the most difficulty is when the camera is moving forward as relatively little parallax is experienced. This results in features taking many observations to become well localized. As we are investigating how inertial measurements can aid in bearing only SLAM this experiment was set up to assess the effect of the inertial measurements in situations that are difficult for normal bearing only implementations.

The experimental data was acquired using raw accelerometer and gyroscope measurements from a MicroStrain 3DM-GX1 inertial measurement unit. Images were taken at a rate of 7.5 frames per second at 640 by 480 pixels using a Logitech QuickCam Pro 4000 with a 90 degree wide angle lens attached. Fig. 3 shows the camera mounted onto the 3DM-GX1 unit. The sensors are connected to a laptop where the data is logged.

Bearing only observations were taken of circular fiducials extracted using SIFT [9]. Data association of the observations were checked manually as it was not the focus of our research.

Fig.4 shows the location of the landmarks used in the experiment. This photo was taken from the initial camera location at the start of the dataset used. Landmarks 0-5 are on the wall at approximately 25 meters from the origin, landmarks 6-9 are on trees approximately 15 meters from the origin and landmarks 9-11 are approximately 6 meters from the origin, where all measurements have been taken by hand.

Fig.5 shows an example of the labeled observation data used by the filter on the right and a plot of the feature map on the left with currently observed features in red (dark),

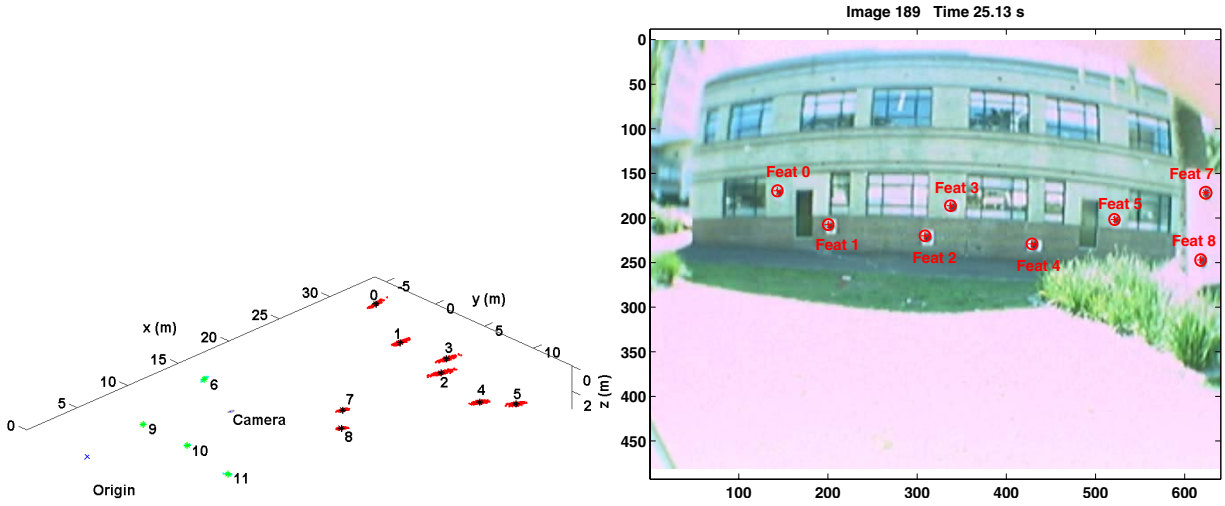


Fig. 5. Screen shot of the filter running showing the visual observations on the right along with ellipses representing the measurement prediction uncertainty and the current map estimate on the left

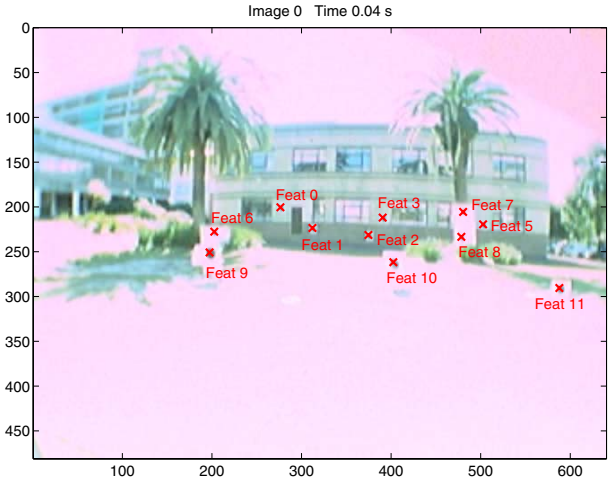


Fig. 4. Photo showing the location of landmarks used in the experiment. Landmarks 0-5 are on the wall at approximately 25 meters from the origin, landmarks 6-9 are on trees approximately 15 meters from the origin and landmarks 9-11 are approximately 6 meters from the origin.

previously observed features in light blue (light) and the camera location uncertainty in dark blue.

IV. Results

The filter was run with and without inertial observations in order to study the influence of the IMU data in the final estimate. We were highly interested in observing if the inertial measurements could aid vision to estimate the scale factor of the map. Another interesting question was the influence of the initial range estimate of the features in the final map.

There are three important issues in the implementation of both methods that have to be taken into account in order to perform a fair comparison.

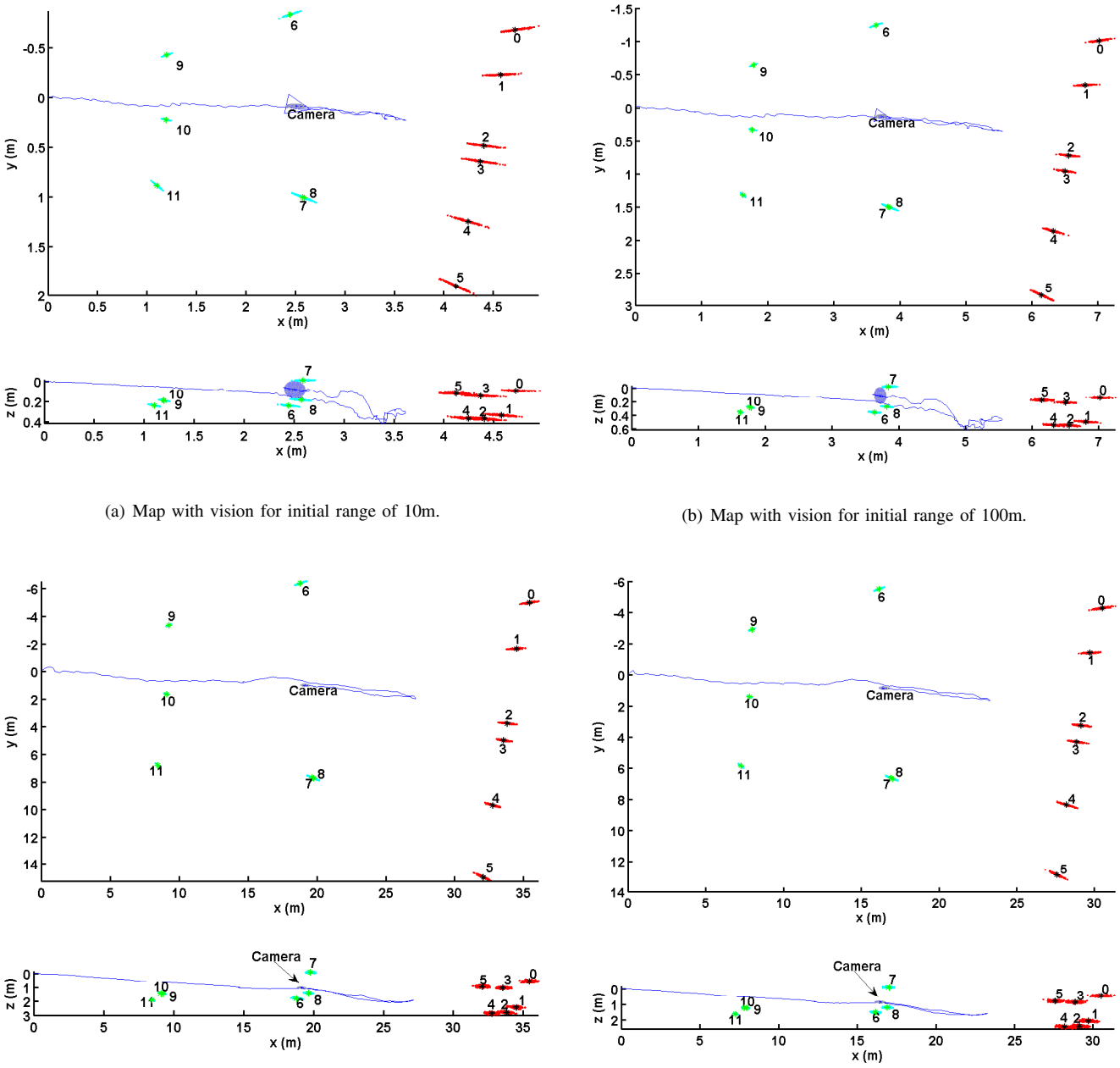
First, since we have less amount of information in the pure monocular SLAM there are slight differences in the prediction step of the filter implementation. The basic change with respect to the inertial case is that the angular velocity ω of the camera has also to be estimated since it is not a system input anymore. Therefore it is included in the state vector. A detailed description of the vehicle state and prediction equations used in the pure monocular SLAM can be found in [5]. There are no more differences in the equations.

Second, a priori information about the initial velocity of the camera is needed in both cases. Otherwise, this value could shift the scale factor making the comparisons impossible. In the experiment the initial velocity was $v_0 = 0m/s$.

Finally, it is of crucial importance in an inertial system to have a rough estimation of the initial attitude. As it is known, attitude and gyros errors (noise and bias) have the most detrimental effect as the gravitational acceleration can not be properly compensated producing significant drifts in the velocity and position estimates. To estimate the initial attitude of the inertial system (angles Roll and Pitch) a coarse alignment method [7] has been used whereas the system was still during the first steps of the experiment.

Results with and without inertial observations and with different initial range estimates are shown in Fig.6. The forward trajectory followed by the camera as well as the distribution of the landmarks represents a singular motion which makes the estimation with bearing only SLAM difficult. As can be observed, initially the camera goes forward until is next to the building wall, going backwards almost along the same path during the final steps of the experiment.

Figs.6(a) and 6(b) show the results obtained with bearing only (without inertial aiding). Features in the left plot are initialized with an initial range estimate of 10 meters whereas features in the right plot are initialized at 100 meters. Although the initial range has changed significantly, one order of magnitude, the influence in the estimation of the map scale is moderate, around 1.5 times bigger. However, what is clear from the figures and from the real distribution of the features (Section III) is that the scale factor is much smaller than it should be. It is approximately 0.175 times the



(a) Map with vision for initial range of 10m.

(b) Map with vision for initial range of 100m.

(c) Map with vision + IMU for initial range of 10m.

(d) Map with vision + IMU for initial range of 100m.

Fig. 6. Comparison of the final maps obtained using monocular SLAM with and without inertial measurements, for different initial range estimates. In all cases the map is represented seen from above (top) and from a lateral perspective (bottom). Landmarks 0-5 are on the wall at approximately 25 meters from the origin, landmarks 6-9 are on trees approximately 15 meters from the origin and landmarks 9-11 are approximately 6 meters from the origin.

real scale for the plot on the left and 0.26 times for the plot on the right.

An interesting behavior of the trajectory can be observed in the lateral perspective figures. The erratic trajectory described by the camera after passing through features 6-8 is due to the singular disposition of the landmarks stuck on the wall. Their vertical baseline does not seem enough to constrain and estimate the vertical trajectory of the camera which explains its strange behavior.

Figs.6(c) and 6(d) show the results obtained with inertial aiding. As in the previous plots, features in the left are initialized at 10 meters and the ones in the right at 100 meters. In this case, the influence of the initial range in the map scale is very small. On the other hand, the estimation of the real scale factor seems much better than before. It is approximately 1.33 times the real scale for the plot on the left and 1.15 times for the plot on the right. Another visible improvement is observed in the trajectory of the camera after passing through features 6-8. With the aid of the inertial the trajectory is much smoother than before which indicates that the IMU is of great help in singular situations where the camera alone has problems.

V. Conclusions

This paper has shown the benefit that a bearing only implementation of SLAM can get from the use of an inertial measurement unit especially when initializing features in situations that are difficult for bearing only sensors such as during forward motion with a forward looking camera. The IMU observations produce a more realistic map by reducing the variation in the estimation of the scale factor of the map.

The inertial data can constrain as well the uncertainties in the prediction of the camera motion between observations which helps with data association and reduce linearisation errors on observations. It has shown to be also helpful during periods of no observations or in singular situations where vision only presents difficulties.

Currently the authors are testing the performance of this filter on a wider range of datasets to evaluate the benefit of inertial observations in different situations. As a extension of this work the effects of the inertial observations on the data association process through providing more accurate camera position predictions is being investigated.

VI. Acknowledgments

This research has been funded in part by the Dirección General de Investigación of Spain under projects DPI2003-07986, DPI2006-13578 and ARC Discovery Grant DP0665439.

Thanks to José M. Montiel and Javier Civera for the fruitful discussions about the inverse depth visual SLAM.

References

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

[2] T. Bailey, "Constrained initialization for bearing only slam," in *IEEE Int. Conf. on Robotics and Automation, ICRA*, Taipei, Taiwan, 2005, pp. 1966–1971.

[3] M. Bryson and S. Sukkarieh, "Bearing only slam for an airborne vehicle," in *Australasian Conference on Robotics and Automation, ACRA*, Sydney, Australia, 2005.

[4] J. Sola, A. Monin, M. Devy, and T.Lemair, "Undelayed initialization in bearing only slam," in *2005 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Edmonton, Canada, 2005, pp. 2499–2504.

[5] J. M. M. Montiel, J. Civera, and J. Davison, "Unified inverse depth parametrization for monocular slam," in *Robotics Science and Systems, RSS*, Philadelphia, Pennsylvania, 2006.

[6] S. Sukkarieh, E. M. Nebot, and H. Durrant-White, "A high integrity IMU/GPS navigation loop for autonomous land vehicle applications," *IEEE Trans. Robot. Automat.*, vol. 15, pp. 572–578, September 1999.

[7] D. H. Titterton and J. L. Weston, *Strapdown Inertial Navigation Technology*. Stevenage, U.K.: Peregrinus, 1997.

[8] J. Kim and S. Sukkarieh, "Airborne simultaneous localisation and map building," in *IEEE Int. Conf. on Robotics and Automation*, Taipei, Taiwan, September 2003.

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