Vision-based SLAM using natural features in indoor environments

[#]Jaime Valls Miró, Gamini Dissanayake, Weizhen Zhou
ARC Centre of Excellence for Autonomous Systems (CAS)
Faculty of Engineering, University of Technology Sydney (UTS)
NSW2007, Australia
{*j.vallsmiro,g.dissanayake,w.zhou*}*@cas.edu.au*

Abstract

This paper presents a practical approach to solve the simultaneous localization and mapping (SLAM) problem for autonomous mobile platforms by using natural visual landmarks obtained from an stereoscopic camera. It is an attempt to depart from traditional sensors such as laser rangefinders in order to gain the many benefits of nature-inspired information-rich 3D vision sensors. Whilst this makes the system fully observable in that the sensor provide enough information (range and bearing) to compute the full 2D estate of the observed landmarks from a single position, it is also true that depth information is difficult to rely on, particularly on measurements beyond a few meters (in fact the full 3D estate is observable, but here robot motion is constrained to 2D and only the 2D problem is considered). The work presented here is an attempt to overcome such a drawback by tackling the problem from a partially measurable SLAM perspective in that only landmark bearing from one of the cameras is employed in the fusion estimation. Range information estimates from the stereo pair is only used during map building in the landmark initialization phase in order to provide a reasonably accurate initial estimate. An additional benefit of the approach presented here lies in the data association aspect of SLAM. The availability of powerful feature extraction algorithms from the vision community, such as SIFT, permits a more flexible SLAM implementation separated from feature representation, extraction and matching, essentially carrying out matching with minimal recourse to geometry. Simulation results on real data illustrate the validity of the approach.

1. Introduction and motivation

The incremental construction of an stochastic map of the environment while concurrently generating an estimate for the location of a mobile vehicle is known as the simultaneous localization and mapping (SLAM) problem. The process can be expressed as follows: starting from an arbitrary initial point, a mobile robot should be able to autonomously explore the environment with its on-board sensors, gain knowledge about it, build an appropriate map and localise itself relative to this map. The resulting model can then be employed by planning and navigational strategies to achieve the robot goal positions in an efficient manner. Whilst accomplishing this goal is still years away, it is a very active area of research and a number of robotics research groups are making substantial contributions in this area (see for example, [1], [2], [3], [4],[5],[6], [7] and the references therein).

Vision SLAM in particular has seen many advances in recent years due to the low cost, light weight and low power requirements of the sensor [8], [9], [10]. They provide a wealth of 3D information from the scene, matched only by few other sensors. It has also the added benefit of enabling roboticist to incorporate advances from the image processing community, in particular approaches to efficiently represent the salient regions of an image (such as the SIFT algorithm employed here), and reliable data association algorithms, active areas of research by the vision research community and an essential component of SLAM.

Monocular sensors can not directly retrieve depth information from the scene. Hence, the traditional Extended Kalman Filter (EKF) technique to solve the SLAM problem can not be readily applied with single cameras [11]. Special landmark initialization techniques have been proposed in the literature to overcome this, thus enabling a full Gaussian estimate of its estate and the application of EKF. This amount to either delayed landmark initialization, where multiple observations are combined from multiple poses [12], or undelayed initialization, where each landmark is initialized in the form of multiple range hypothesis, and subsequent measurements will dictate the one to remain on the map, while the others are removed [13]. Stereo vision systems on the other hand can provide depth information from the surrounding area. The disparity map provided by a stereo system can be used to determine 3D coordinates to a point features in the environment. This has been used in [10], although the approach is not globally consistent as the cross-correlations where simplified and not fully maintained for computational reasons. Other approaches rely on iterative minimization algorithms from vision techniques such as ICP to perform local 3D alignments to solve the SLAM problem [14]. However, camera calibration and stereo correlation in general are not robust or reliable enough to provide accurate depth maps within the sensor ranges, ready to be used in classic EKF.

In this paper we propose a simple solution to the problem in the form of a bearing-only SLAM implementation with undelayed landmark initialisation. This is accomplished by

assuming the initial location of a landmark as that provided by the stereo observation. Simulation results have shown that as long as the initial estimate is a good one, EKF can deal with the uncertainties associated [13]. It is to that purpose that the availability of two simultaneous bearing observations can be used. After landmark initialisation, the process is that of a bearing-only EKF implementation when only visual observations from one of the cameras is incorporated into the filter. It is worth pointing out that the approach is readily applicable to platforms with two (or more) general-purpose camera sets, not necessarily proper epipolar stereoscopic headsets, since the critical component is that of having more than one simultaneous observation to the feature to ascertain the required depth information. Calibration and set up of non-epipolar cameras is however even more complex and difficult to get right that stereo heads, and the work presented here makes use of a commercial off-the-shelf stereo headset.

The work presented here is currently being extended by the authors to make better use of the full range and bearing information from the pair of cameras, in an attempt to implement a practical, robust, full range and bearing multicamera-based SLAM implementation with invariant features extracted from the scene.

The remainder of this paper is organised as follows: Section 2 summarises the mathematical framework employed in the study of the SLAM problem. Section 3 reviews the relevant aspects of the SIFT algorithm, with a discussion of data association issues in SLAM. Then, the proposed methodology for the solution to the visual-based SLAM problem is presented in Section 4. Detailed experimental setup and results with real stereo data and robot odometry are provided in Sections 5 and 6 respectively. Finally, Section 7 summarises the contribution of this paper.

2. THE SLAM PROBLEM FORMULATION

This framework is provided here for completeness and to facilitate the discussion in the later sections. An interested reader is referred to [1] for a more comprehensive description.

A. Vehicle and landmark models

The setting for the SLAM problem is that of a vehicle with a known kinematic model, starting at an unknown location, moving through an environment containing a population of features or landmarks. The vehicle is equipped with a sensor that can take measurements of the relative location between any individual landmark and the vehicle itself as shown in Figure 1. The state of the system consists on the position and orientation of the vehicle together with the position of all landmarks. The state of the vehicle at a time step k is denoted $\mathbf{x}_v(k)$. The motion of the vehicle through the environment is modelled by a conventional linear discrete-time state transition equation or process model of the form

$$\mathbf{x}_v(k+1) = \mathbf{F}_v(k)\mathbf{x}_v(k) + \mathbf{u}_v(k+1) + \mathbf{v}_v(k+1), \quad (1)$$

where $\mathbf{F}_v(k)$ is the state transition matrix, $\mathbf{u}_v(k)$ a vector of control inputs, and $\mathbf{v}_v(k)$ a vector of temporally uncorrelated process noise errors with zero mean and covariance $\mathbf{Q}_v(k)$. The location of the i^{th} landmark is denoted $\mathbf{p}_i$. The "state transition equation" for the i^{th} landmark is

$$\mathbf{p}_i(k+1) = \mathbf{p}_i(k) = \mathbf{p}_i, \quad (2)$$

since landmarks are assumed stationary. The vector of all N landmarks is denoted as

$$\mathbf{p} = \begin{bmatrix} \mathbf{p}_1^T & \dots & \mathbf{p}_N^T \end{bmatrix}^T \quad (3)$$

The augmented state vector containing both the state of the vehicle and the state of all landmark locations is denoted

$$\mathbf{x}(k) = \begin{bmatrix} \mathbf{x}_v^T(k) & \mathbf{p}_1^T & \dots & \mathbf{p}_N^T \end{bmatrix}^T. \quad (4)$$

The augmented state transition model for the complete system may now be written as

$$\begin{bmatrix} \mathbf{x}_v(k+1) \\ \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_N \end{bmatrix} = \begin{bmatrix} \mathbf{F}_v(k) & 0 & \dots & 0 \\ 0 & \mathbf{I}_{p_1} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{I}_{p_N} \end{bmatrix} \begin{bmatrix} \mathbf{x}_v(k) \\ \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_N \end{bmatrix} + \begin{bmatrix} \mathbf{u}_v(k+1) \\ \mathbf{0}_{p_1} \\ \vdots \\ \mathbf{0}_{p_N} \end{bmatrix} + \begin{bmatrix} \mathbf{v}_v(k+1) \\ \mathbf{0}_{p_1} \\ \vdots \\ \mathbf{0}_{p_N} \end{bmatrix} \quad (5)$$

$$\mathbf{x}(k+1) = \mathbf{F}(k)\mathbf{x}(k) + \mathbf{u}(k+1) + \mathbf{v}(k+1) \quad (6)$$

where $\mathbf{I}_{p_i}$ is the $\dim(p_i) \times \dim(p_i)$ identity matrix and $\mathbf{0}_{p_i}$ is the $\dim(p_i) \times \dim(p_i)$ null matrix.

B. The observation model

The vehicle is equipped with a sensor that can obtain observations of the relative location of landmarks with respect

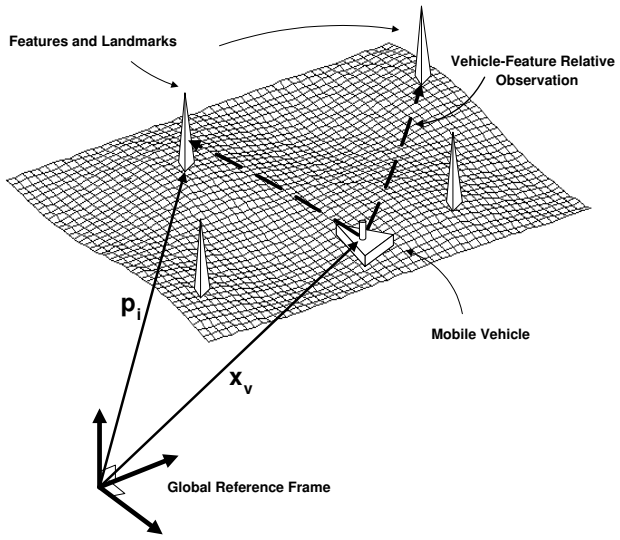

Fig. 1: The SLAM problem: a vehicle takes relative observations of environment landmarks. The absolute location of landmarks and vehicle are unknown.

to the vehicle. The observation model for the i^{th} landmark is assumed to be linear and synchronous, and written in the form

$$\begin{aligned} \mathbf{z}_i(k) &= \mathbf{H}_i\mathbf{x}(k) + \mathbf{w}_i(k) \\ &= \mathbf{H}_{p_i}\mathbf{p} - \mathbf{H}_v\mathbf{x}_v(k) + \mathbf{w}_i(k) \end{aligned} \quad (7)$$

where $\mathbf{w}_i(k)$ is a vector of temporally uncorrelated observation errors with zero mean and variance $\mathbf{R}_i(k)$. It is important to note that the observation model for the i^{th} landmark is written in the form

$$\begin{aligned} \mathbf{H}_i &= [-\mathbf{H}_v, \mathbf{0}\cdots\mathbf{0}, \mathbf{H}_{p_i}, \mathbf{0}\cdots\mathbf{0}] \quad (8) \\ &= [-\mathbf{H}_v, \mathbf{H}_{mi}] \quad (9) \end{aligned}$$

This structure reflects the fact that the observations are taken from on-board the vehicle and are therefore a measurement of some relative relationship between the vehicle and the feature. This relative relationship is often in the form of relative relative range and bearing, or only bearing like in the work presented here.

C. The estimation process

The Kalman filter is the sensor fusion technique used in SLAM to provide estimates of vehicle and landmark location. The Kalman filter recursively computes estimates for a state $\mathbf{x}(k)$ which is evolving according to the process model in Equation 6 and which is being observed according to the observation model in Equation 7. The Kalman filter computes an estimate which is equivalent to the conditional mean $\hat{\mathbf{x}}(p|q) = E[\mathbf{x}(p)|\mathbf{Z}^q]$ ($p \geq q$), where $\mathbf{Z}^q$ is the sequence of observations taken up until time q. The error in the estimate is denoted $\tilde{\mathbf{x}}(p|q) = \hat{\mathbf{x}}(p|q) - \mathbf{x}(p)$. The Kalman filter also provides a recursive estimate of the covariance $\mathbf{P}(p|q) = E\left[\tilde{\mathbf{x}}(p|q)\tilde{\mathbf{x}}(p|q)^T|\mathbf{Z}^q\right]$ in the estimate $\hat{\mathbf{x}}(p|q)$. The Kalman filter algorithm now proceeds recursively in three stages:

1) Prediction: Given that the models described in equations 6 and 7 hold, and that an estimate $\hat{\mathbf{x}}(k|k)$ of the state $\mathbf{x}(k)$ at time k together with an estimate of the covariance $\mathbf{P}(k|k)$ exist, the algorithm first generates a prediction for the state estimate, the observation (relative to the i^{th} landmark) and the state estimate covariance at time $k+1$ according to

$$\begin{aligned} \hat{\mathbf{x}}(k+1|k) &= \mathbf{F}(k)\hat{\mathbf{x}}(k|k) + \mathbf{u}(k) \quad (10) \\ \hat{\mathbf{z}}_i(k+1|k) &= \mathbf{H}_i(k)\hat{\mathbf{x}}(k+1|k) \quad (11) \\ \mathbf{P}(k+1|k) &= \mathbf{F}(k)\mathbf{P}(k|k)\mathbf{F}^T(k) + \mathbf{Q}(k), \quad (12) \end{aligned}$$

respectively.

2) Observation: Following the prediction, an observation $\mathbf{z}_i(k+1)$ of the i^{th} landmark of the true state $\mathbf{x}(k+1)$ is made according to Equation 7. Assuming correct landmark association, an innovation is generally calculated as follows

$$\nu_i(k+1) = \mathbf{z}_i(k+1) - \hat{\mathbf{z}}_i(k+1|k) \quad (13)$$

together with an associated innovation covariance matrix given by

$$\mathbf{S}_i(k+1) = \mathbf{H}_i(k)\mathbf{P}(k+1|k)\mathbf{H}_i^T(k) + \mathbf{R}_i(k+1). \quad (14)$$

3) Update: The state estimate and corresponding state estimate covariance are then updated according to:

$$\hat{\mathbf{x}}(k+1|k+1) = \hat{\mathbf{x}}(k+1|k) + \mathbf{W}_i(k+1)\nu_i(k+1) \quad (15)$$

and

$$\mathbf{P}(k+1|k+1) = \mathbf{P}(k+1|k) - \mathbf{W}_i(k+1)\mathbf{S}_i(k+1)\mathbf{W}_i^T(k+1) \quad (16)$$

Where the gain matrix $\mathbf{W}_i(k+1)$ is given by

$$\mathbf{W}_i(k+1) = \mathbf{P}(k+1|k)\mathbf{H}_i^T(k)\mathbf{S}_i^{-1}(k+1) \quad (17)$$

The update of the state estimate covariance matrix is of paramount importance to the SLAM problem. It is shown in [1] that as the vehicle progresses through the environment the errors in the estimates of any pair of landmarks become more and more correlated. Furthermore, the entire structure of the SLAM problem critically depends on maintaining complete knowledge of the cross correlation between landmark estimates. The stochastic map of correlations must be maintained and updated as the robot is performing automatic map-building. Minimizing or ignoring cross correlations is precisely contrary to the structure of the problem.

3. VISUAL FEATURE PROCESSING

In the work proposed here an efficient mechanism to detect and represent stable local features was required. An immensely popular choice drawn from computer vision as a fundamental component of many image registration and object recognition is SIFT [10], [15]. Whilst not the only one, a recent comparative study [16] of several local descriptors showed that the ranking of accuracy for different algorithms was relatively insensitive to the method employed to find the interesting points in the image, but was dependant on the representation used to model the image patch around the salient points. Their best matching results were obtained using the SIFT mechanism, which was identified as the most resistant to common image deformations. This made it the sensible choice for our research, and the work of other researchers working on SLAM also seem to agree with this judgement (see, for instance [8] and [10]).

A. Landmark identification and representation: SIFT

To summarise the approach, SIFT consists of four major stages:

- scale-space peak selection, where potential interest points are selected by scanning the image over location and scale. This is accomplished by constructing a Gaussian pyramid and searching for local peaks (keypoints) in a series of difference-of-Gaussian (DoG) images, which makes for an efficient implementation.
- keypoint localization, when candidate keypoints are localized to sub-pixel accuracy and eliminated if found to be unstable.
- orientation assignment, where the dominant orientation for each keypoint based on its local image patch is identified. This is very relevant to SLAM since the assigned

orientation, scale and location for each keypoint enables SIFT to construct a canonical view for the keypoints that is invariant to similarity transformations. From a data association point of view, this is very interesting as it makes image features robust to changes in view point, enabling re-detection of a previously visited area, i.e., closing the loop.

- keypoint descriptor, the final stage where the descriptor for each keypoint is built based upon the image gradients in its local neighbourhood. Magnitudes and orientations of the image gradient in the patch around the keypoint are sampled, and a smooth 4x4 array of orientation histograms is built over them. The 16 cells in the grid, each at 45 degree intervals, captures the rough spatial structure of the patch. This yields a 4x4x8=128 dimensional descriptor vector for each processed region, which is the normalized to unit length and thresholded to remove elements with small values.

The end result is a compact descriptor that allows quick comparisons with other regions, and rich enough to allow these comparisons to be highly discriminatory. This is particularly so as the descriptor representation is designed to avoid problems due to boundary effects, i.e., smooth changes in location, orientation and scale do not cause radical changes in the feature vector. Furthermore, while the representation was not designed to be explicitly invariant to affine transformations, it is nevertheless surprisingly resilient to deformations such as those caused by perspective effects [16].

B. Data association with SIFT

All the above characteristics are evidenced in excellent matching performances, which make them an ideal candidate to the on-going problem in SLAM of robust data association. In particular when the pose estimate of the vehicle is in gross error, which means that despite the fact the vehicle might be in an area already mapped, loop closure with the traditional geometry-based nearest neighbour gating is not detected, resulting in wrong re-mapping and erroneous global locations [8]. Figure 2 shows the relative insensitivity of SIFT to changes in viewpoint from the same scene by correctly matching corresponding keypoints. Lines are not parallel as SIFT points of interest are not found on a planar surface. The good matching characteristics of SIFT descriptors is also applicable to image pairs obtained from the stereoscopic sensor, as seen in Figure 3. This can be used in SLAM to eliminate features with spurious existence, and those which don't lie in the camera epipolar plane, and only surviving keypoints that appear in both left and right images are then allowed to be initialized and integrated in to the SLAM feature database.

4. The Visual-SLAM approach

The algorithm for visual-SLAM as proposed here can be described as follows:

1) Initialization: set up a world coordinate frame at the initial robot location.

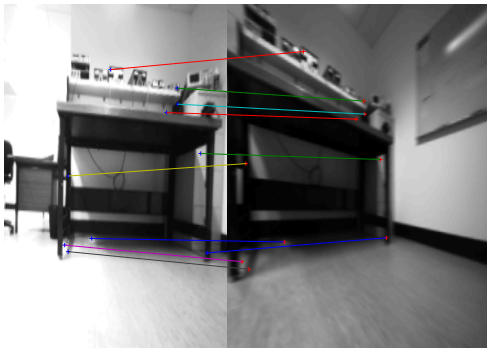

Fig. 2: Corresponding keypoints which show the robustness of SIFT to changes in view point.

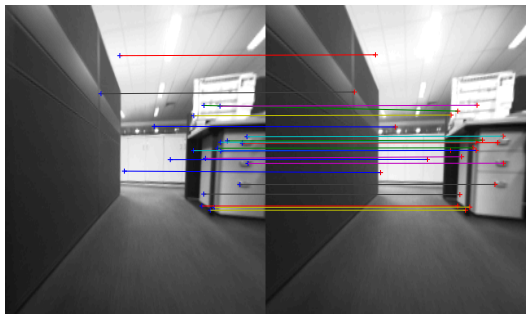

Fig. 3: Matched keypoints between a stereo pair of images.

2) Initialization: obtain a stereo image pair from the scene and run SIFT on both left and right images.
3) Initialization: taking the left as the reference image, find matches by looking for the descriptor vector in the right image with closest Euclidean distance. Some further thresholding is carried out as suggested in [10] to only keep most unique and distinctive features, discarding the feature if it is considered to be too similar to more than one keypoint. Potential mismatches are further filtered out by enforcing keypoints to remain in epipolar planes (this is the only step not applicable should cameras be on different planes).
4) Initialization: calculate the corresponding disparity and depth information for the matched descriptors. Although the validated features are in 3D coordinates, the robot movement is considered to be two-dimensional. This is a reasonable assumption for indoor environments as a typical indoor office-like environment can be safely regarded as flat. Hence, an initial estimate for the 2D location of the local observed feature relative to the robot is attained and included in the initial map of the environment.
5) Loop: repeat Step 2
6) Loop: repeat Step 3

7) Loop: EKF update of robot localization and mapping as explained in Section 2. Robot prediction and predicted observation are carried out based on bearing to features only, ignoring poor quality range information at this stage, with the consequent reduction in computational expense. Features are matched in a similar fashion to Step 3 above taking advantage of the data association characteristics of SIFT, as described in Section 3-B, except epipolar restrictions make no sense here and are removed. A 2σ bearing innovation gate is then implemented for further validation, but results show that a simpler geometric thresholding on the orientation error provides equally good results. In EKF, the new estate estimate must be accompanied by the increase in the state uncertainty (process noise covariance) for the camera after this motion. Jacobians (first partial derivatives) are employed for that. While tractable, this cross-correlation updates are costly, so only those features which have been re-observed a number of times are employed to carry out the estate Jacobian update.
8) Loop: some form of map maintenance needs to be implemented, as the number of observed features can grow very large and tracking them all can become computationally very expensive. A two-fold process has been devised to only retain the most significant point features: firstly, only after a predetermined number of frames (5 to 10 based on experimental results) are re-observed beacons regarded as a permanent features and added to the map. Secondly, beacons which have been matched a small number of times are pruned out of the state vector together with the corresponding entries in the system covariance matrix. How often this is carried out depends mostly on the processing power and memory available, as well as the run itself: longer runs need to prune more often to reduce the map and the expense of tracking a large number of landmarks. In our experience, every 20 to 30 frames was sufficient to produce a manageable map. These two values, as well as the number of successful hits to a feature before it is included in the map (3 to 5 in our simulations), are arbitrary and at the moment based on a trade-off between map density and accuracy, and computational complexity. It is envisaged this will become a more relevant issue when the algorithm is implemented in the real platform, and active steps are currently being investigated to learn how to automatically adjust these parameters on the run.
9) Loop: go back to Step 5.

5. EXPERIMENTAL SETUP

To test the validity of the approach data was collected with an ActiveMedia Pioneer 2DX robot mounted with an stereoscopic camera, and a laser range finder for reference (see Section 5-B below). The robot was driven through an office environment while capturing real-life stereo images, odometric poses and laser scan measurements at around 4Hz. Stereo and pose logged data was then simulated by batch-processing the data through the algorithm described earlier in Section 4, running the Matlab. Further verification is being conducted by running the algorithms in the robot in real-time, although this is still work in progress.

A. Stereoscopic headset

The stereo head is the STH-MDCS from Videre Design, a compact, low-power colour digital stereo head with an IEEE 1394 digital interface. It consists of two 1.3 megapixel, progressive scan CMOS imagers mounted in a rigid body, and a 1394 peripheral interface module, joined in an integral unit. Wide-angle lenses (around 100 degrees field of vision) were fitted for this exercise (narrow angle lenses were also tested with poorer results as a lesser number of good quality distinctive features were picked up). The camera was mounted at the front and top of the vehicle at a constant orientation, looking forward. Images obtained were restricted to greyscale $320x240$ pixels, although the sensor can do larger sizes.

B. Laser for performance comparison

The robot was also equipped with a SICK LMS200 laser rangefinder. The range and bearing measurements was used separately to compute a 2D SLAM algorithm for ground truth comparison and validation purposes. A more accurate localizer is currently being implemented with positive results and these will also be included in the final manuscript.

C. Software environment

The widely used Player open source robotics architecture, running under Linux, was the software of choice to interface with the robotic platform and the sensors to perform the synchronous data collection and actual control of the robot. The SRI Small Vision System (SVS) software was employed to calibrate the stereo head and perform stereo correlation within the Player framework.

6. RESULTS

The experimental workspace is that of a classical office open space, with around 1.5 m. height partitions and narrow corridors, as depicted by Figure 3. An example is shown where the robot is driven around two adjacent partitions in an area of around 6x10 m., closing the loop on two occasions as is clearly visible from the state covariances in Figure 5. Results from SLAM are shown in Figure 4 with a reference to (a laser EKF) ground truth, where it is apparent that whilst the robot path sequence still exhibits some error, it is nevertheless a vast improvement over the robot dead reckoning output. As maps are not post-processed, and features are projected in 2D, those in the ceiling often appear in open spaces. The authors are aware this scenario is not overly challenging for current laser-based SLAM solutions, yet the built-up spatial error accumulated over time is shown to be already important and serves to demonstrate the effectiveness of the approach.

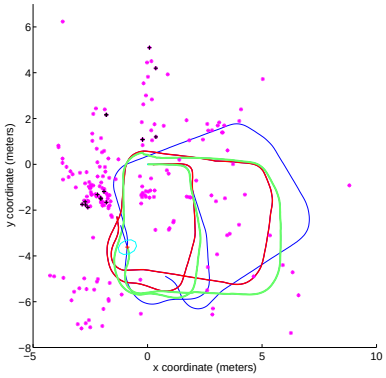

Fig. 4: Robot odometric path (blue, thin), ground truth (red, thick) and SLAM robot pose estimate (green, thicker). Final map landmarks are shown in magenta stars

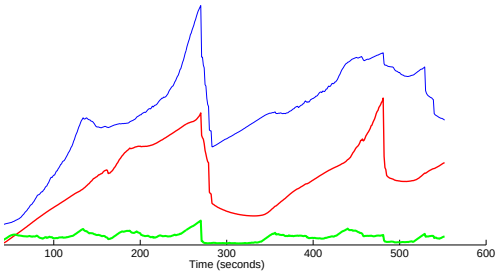

Fig. 5: SLAM estimate covariances: [x (blue, thin), y (red, thick), θ (green, thicker)]

7. Conclusions and Further Work

A solution to the SLAM problem with vision sensors in an unmodified indoor environment has been proposed using an extended Kalman filter. The approach presented here assumes the availability of simultaneous visual information from two cameras. However, only landmark bearing information from one of the camera is utilised in the filtering process. It is suggested that basic range information from the stereo pair is employed, along with bearing, only to provide the initial hypothesis of the new feature (2D) location to be estimated by the filter framework. This reduces filter complexity, particularly with a view to implementation in a real robot, but what is more relevant, eliminates the uncertainty introduced by stereo camera calibration, very sensitive to changes in the scene such as lighting conditions. Powerful data association based on SIFT descriptors, as proposed here, allows for a simple geometric gating on bearing error to enable loop closure even when feature density is high and nearest neighbour data association on its own is impractical. A map management to eliminate less informative landmarks has also been implemented. Further work is currently being undertaken to implement the approach in a real robot in real time. Initial encouraging results show the viability of the approach. Although experimental tests so far have been limited to smaller indoor area, there is a motivation to extend these to more challenging data sets. Further research is also concentrated on augmenting the filter to also incorporate range estimates from the stereo set into the filter. It will be interesting to compare results between the bearing-only implementation proposed here and a full stereoscopic-based range-and-bearing SLAM with the robust data association exhibited by SIFT or similar visual descriptors.

Acknowledgements

This work is supported by the Australian Research Council (ARC) through its Centre of Excellence programme, and by the New South Wales State Government. The ARC Centre of Excellence for Autonomous Systems (CAS) is a partnership between the University of Technology Sydney, the University of Sydney and the University of New South Wales.

References

[1] G. Dissanayake, P. Newman, S. Clark, H. Durrant-Whyte, M. Csorba, "A solution to the simultaneous localization and map building (SLAM) problem", *IEEE Trans. on Robotics and Automation*, vol. 17, pp. 229–241, 2001.

[2] J. E. Guivant, E. M. Nebot, "Optimization of the simultaneous localization and map building (SLAM) algorithm for real time implementation", *IEEE Trans. on Robotics and Automation*, vol. 17, pp. 242–257, 2001.

[3] P. Newman. On the Structure and Solution of the Simultaneous Localization and Map Building Problem. PhD thesis, Australian Centre of Field Robotics, University of Sydney, Sydney, 2000.

[4] J. A Castellanos, J. Neira, J. D. Tardos, "Multisensor fusion for simultaneous localization and map building", *IEEE Trans. on Robotics and Automation* vol. 17, 908–914, 2001.

[5] J. Leonard, P. Newman, "Consist ant, convergent and constant time SLAM", *International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 2003, pp. 1143–1150.

[6] J. Folkesson, H. I. Christensen, "Graphical SLAM - a self-correcting map", *Proceedings IEEE International Conference on Robotics and Automation*, New Orleans, LA, pp. 383-390, 2004.

[7] S. Thrun, Y. Liu, D. Koller, A. Y. Ng, Z. Ghahramani, H. Durrant-Whyte, "Simultaneous localization and mapping with sparse extended information filters", *International J. of Robotics Research*, vol. 23, pp. 693–716, 2004.

[8] P. Newman, K. Ho, "SLAM - loop closing with visually salient features", in *Proc. IEEE Int. Conference on Robotics and Automation*, Barcelona, Spain, pp. 644–651, 2005.

[9] A. J. Davison, D. Murray, "Simultaneous localization and map building using active vision", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2002.

[10] D. G. Lowe, S. Se, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks ", *International Journal of Robotics Research, 21(8)*, pp. 735–758, 2002.

[11] T. Bailey, "Constrained initialization for bearing-only SLAM", *Proc. IEEE Int. Conf. on Robotics and Automation*, Taipei, Taiwan, September 2005.

[12] P. Newman, J. Leonard, R. Rikoski, M. Bosse, "Mapping partially observable features from multiple uncertain vantage points", *Int. J. of Robotics Research*, January 2002.

[13] N. M. Kwok, G. Dissanayake, "An efficient multiple hypothesis filter for bearing-only SLAM", *IEEE Int. Conf. on Intelligent Robot Systems (IROS)*, 2004.

[14] J. M. Saez, F. Escolano, "Entropy minimization SLAM using stereo vision", in *Proc. IEEE Int. Conf. on Robotics and Automation*, Barcelona, Spain, pp. 36–43, 2005.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol. 60(2), pp. 91–110, 2004.

[16] , K. Mikolajczyk, C. Schmid, "A performance evaluation of local descriptors", in *Proc. of IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, Madison, USA, pp. 257-263, 2003.