Information-Driven 6D SLAM Based on Ranging Vision

Weizhen Zhou, Jaime Valls Miró and Gamini Dissanayake
ARC Centre of Excellence for Autonomous Systems (CAS)
Faculty of Engineering, University of Technology, Sydney (UTS)
NSW 2007, Australia
{*w.zhou,j.vallsmiro,g.dissanayake*}*@cas.edu.au*

Abstract— **This paper presents a novel solution for building three-dimensional dense maps in unknown and unstructured environment with reduced computational costs. This is achieved by giving the robot the 'intelligence' to select, out of the steadily collected data, the maximally informative observations to be used in the estimation of the robot location and its surroundings. We show that, although the actual evaluation of information gain for each frame introduces an additional computational cost, the overall efficiency is significantly increased by keeping the matrix compact. The noticeable advantage of this strategy is that the continuously gathered data is not heuristically segmented prior to be input to the filter. Quite the opposite, the scheme lends itself to be statistically optimal and is capable of handling large data sets collected at realistic sampling rates. The strategy is generic to any 3D feature-based simultaneous localization and mapping (SLAM) algorithm in the information form, but in the work presented here it is closely coupled to a proposed novel appearance-based sensory package. It consists of a conventional camera and a range imager, which provide range, bearing and elevation inputs to visual salient features as commonly used by three-dimensional point-based SLAM, but it is also particularly well adapted for lightweight mobile platforms such as those commonly employed for Urban Search and Rescue (USAR), chosen here to demonstrate the excellences of the proposed strategy.**

I. INTRODUCTION

One of the many important applications of mobile robots is to reach and explore terrains which are inaccessible or considered dangerous to humans. Such environments are frequently encountered in USAR scenarios where, prior knowledge of the environment is unknown but required before any rescue operation can be deployed. A small mobile robot equipped with an appropriate sensor package can possibly be regarded as the best aid in such scenario. The robot is expected to navigates itself through the site and generate maps of the environment which human rescuers can then use for navigating and locating victims. Despite the wealth of research in planar robot mapping, the limitations of traditional 2D feature maps in providing useful and understandable information in these scenarios have propelled increasing research efforts towards the generation of rich textural 3D maps instead.

3D mapping of a USAR scenario using a light-weight and highly mobile autonomous robot is a challenging predicament which can be framed within the generic simultaneous localization and mapping (SLAM) problem [4], where a robot is expected to move with six degrees-of-freedom in a three-dimensional environment. This, per se, demanding picture is further complicated by the lack of odometry information from the wheel encoders, as this tends to be totally unreliable due to the nature of the disaster environment. The emerging approach based on the ubiquitous laser range scanner lies in deriving odometry from 3D scan matching, possibly fusing textured images a-posteriori to recover the final shape [9]. When the noise on the relative locations is small, as is the case of accurate range sensor such as tilting laser range finders or other forms of egomotion sensors, such as large and expensive IMUs, a good quality estimate of the final 3D map can be expected if displacement between robot poses is limited. By doing so the SLAM problem is reduced to that of the estimation of a state vector containing all the camera/robot poses. This is the approach used by many researchers in the SLAM community, e.g. [3], [11], [7].

Wide availability of lower cost, lower power, lighter-weight passive cameras as well as maturity of computer vision algorithms have made real-time vision processing much more practical in recent times, and consequently there has been an increasing interest in visually based navigation systems in the robotics community. Cameras are particularly interesting as they provide a wealth of geometric information from an unmodified scene, as well as perceptual information such as textures and colours, which can be matched by few other sensors. Vision SLAM in particular has seen many advances in recent years [10], and our efforts have also veered towards sequential 3D dense map reconstruction within a visual SLAM framework to meet the challenges intrinsic to disaster environments such as USAR.

In previous research work [12] an alternative compact visual sensor package along with a SLAM algorithm has been proposed. A traditional passive pin-hole camera is combined with one of the recently developed low resolution range imagers. The conventional camera is used to capture scene texture and to extract salient visual features whilst the range camera provides 3D data of the corresponding scene. The combined observations made by these two cameras are then used as the sole input in the SLAM process thereafter. The technique employs a conventional EIF approach which recovers the robot and feature poses at the end of each 'acquire, update' cycle. However, the inversion of information matrix evoked during each estimation cycle comes at a significant computational cost. Although the sensor package is capable of delivering data at a minimal frame rate of 5 to 10 Hz, which should theoretically guarantee in itself appropriate frame registration, the increasing sampling delay between consecutive frames caused by extensive computation yields inadequate data association and therefore unsuccessful frame registration. Such problem is particularly magnified in unstructured environment where the robot's sights change rapidly and unpredictably along its undulating path.

Many efforts have been made in recent years to reduce the computational encumbrance generally faced by most SLAM algorithms, particularly in its most efficient information form. In related work [8], Eustice *et al* implemented an Exactly Sparse Delayed-State Filter (ESDSF) which maintained a sequence of delayed robot poses at places where low-overlap images were captured manually. In [11], the authors also used a Delayed State Extended Kalman Filter (EKF) to fuse the data acquired with a 3D laser range finder. A segmentation algorithm was employed to separate the data stream, based on orientation restrains, into distinct point clouds, each referenced to a vehicle position. Both implementations significantly reduced the computational cost by eliminating features from the state vector to achieve practical performance. However, one noticeable common problem of these strategies is that loop closure can not be automatically detected. Separate loop closure methods were required in conjunction with their proposed techniques. Furthermore, both methods require either human supervision over the data acquisition process or raw odometry measurements to minimize the number of critical robot poses that should be maintained, none of which are available to us in the settings of a USAR scenario.

It was demonstrated [12] that dense 3D maps can be constructed with carefully prepared data sets collected in a static fashion, in the sense that each camera pose was manually situated to ensure extensive overlapping between consecutive frames as well as the full coverage of the arena. However, in real applications such as USAR, it is unrealistic to expect such 'ideal' data sets. The sensor package is more likely to be operated at its maximum rate to overcome problems such as motion blurriness and drastic changes in the scene due to the inherent nature of unstructured environments. Thus an alternative approach to address the computational issues encountered in more realistic settings is proposed here: instead of focusing on minimising the information gathered and trying to compute them in mathematically efficient ways, we seek a solution where we can collect information at maximum sensor rates and give the robot the 'intelligence' to choose the critical observations that should be incorporated in the estimation process. To accomplish that, in this paper we extend our current methodology with an improved filtering technique whereby given a desired estimation error boundary, a buffer of overlapped continuous visual scans are sampled but only those providing maximal information gain are actually introduced in the filter. With this technique, the filter incorporates only a minimal number of robot poses, but critically distributed along the trajectory in an automatic manner based on the robot uncertainty belief. Moreover, we can also afford to maintain both robot and feature poses in the state vector, which provides a more accurate estimation over camera poses and automates loop closure.

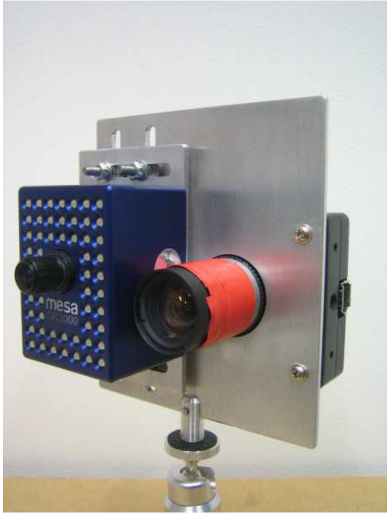


Fig. 1. The sensor package: a conventional pinhole camera aligned with the SwissRanger SR-3000 ranger

The consistency of the proposed strategy has been examined and validated with simulated data in [13]. In this paper, we present results obtained from real data sets collected by the proposed sensor package while operating at a realistic sampling rate. The outcome of a reconstructed 3D dense map that closely reassembles the environment being explored alongside pictures from the scene are presented to show the qualities of the strategy.

The rest of this paper is structured as follows: Section II describes the visual sensor package. The 3D feature extraction and registration process is explained in Section III. Section IV covers the mathematical formulation of the EIF SLAM algorithm and the proposed information-efficient strategy. Section V presents the experimental results. Discussion and concluding remarks are drawn in VI where improvements and future work directions are also proposed.

II. The Visual 3D Sensor Package

In this work, we have employed an improved version of the sensor package used in [12] which consists of a time-of-flight range camera (SwissRanger SR-3000, low resolution, 176×144 pixels) and a higher resolution conventional camera (Point Grey Dragonfly2, 1024×768 pixels). The two cameras are fixed relative to each other as illustrated in Fig. I.

The SwissRanger works on the principle of emitting modulated infra-red light on the scene with a 20 MHz frequency and then measuring the phase shift of the reflection to provide 3D range data without an additional tilting/panning mechanism, albeit within a limited range (the known non-ambiguity range of the sensor [5] is 7.5 meters). Further to distance information, the SwissRanger is also able to return information about the reflected signal's amplitude, hence capturing an intensity image of the scene. However, this is currently too noisy and subject to substantial changes in illumination as the camera pose changes. Hence the proposal for a conventional camera, insensitive to the infra-red light emitted by the SwissRanger, to capture scene texture and to extract salient visual features to aid the SLAM algorithm.

III. Feature Extraction and Frame Registration

With no prior knowledge of the robot motion nor the scene, an efficient mechanism to estimate the relative pose between two images is required as an input to the filter. A popular choice drawn from computer vision as a fundamental component of many image registration and object recognition algorithms is the Scale Invariant Feature Transformation (SIFT) [6]. The main strength of SIFT is to produce a feature descriptor that allows quick comparisons with other features, and is rich enough to allow these comparisons to be highly discriminatory.

The process for the frame registrations is as follows: firstly the two cameras are stereo calibrated. SIFT features are then detected in the 2D camera image and matched across those in the previous images. Due to the offset between the two cameras, there is not a one-to-one corresponding pixel which can be obtained directly from the SwissRanger's intensity image. However, given the calibration information, we can compute the 3D position at where the feature visual cue (bearing) should intersect the 3D point cloud. If a point can be located around the intersection point and its distance is within the known SwissRanger's measurement precision at that depth, we register this 3D point as a potential feature as illustrated in Fig. 2 for a single SIFT point. Applying a least square 3D point set registration algorithm [2] and an outlier removal [1], we obtained a subset of features which we can use for estimating the initial value of the new camera pose with the previous camera pose as prior.

IV. Efficient Extended Information Filter SLAM

A. Information Efficient Filtering Strategy

As described in the introduction section, in USAR scenario the desire is for a system which can deliver not only maximal information but also a human comprehensible presentation of the environment in minimal time. To do so, a "look-ahead and search backwards" algorithm is proposed. The idea is maximising data collection by buffering up all the information available but only choosing the most crucial data to be processed for mapping and localisation. Assuming the robot begins at the origin $[0, 0, 0]$ of the global coordinate frame at time $t = 0$, the feature global poses can be established and be used as the first 'base' frame, F_{base}. For the following frames, matching features are found between F_{base} and each individual frame, unless the number of common features reaches a predefined minimum or the number of frames being examined exceeds the look-ahead buffer size. The minimal number of common features is restricted by the 3D registration algorithm [2] to 6, while buffer size is an empirical number determined by the desired coarseness of the map. 3D registration is performed on each frame with respect to their matching F_{base} (all global coordinates). Given new observations made at each new robot pose, information gain can be obtained without recovering the state vector which is the major computational expense in EIF SLAM. The camera pose at which the observations provide maximal information gain is added to the filter, and the corresponding frame is included as a new entry in the F_{base} database. Frames in the look-ahead buffer previous to the new F_{base} are dropped, and a new buffer is started from the consecutive frame after the last entry in F_{base}. The same procedure is repeated until the end of the trajectory. The reader is referred to [13] for a more detailed discussion of this process.

Although the robot is not processing every frame it acquired during the traversed course, its knowledge is not limited to a set of known positions as is also in possession of additional non-filtered information which it can retrieve to increase the chances to regain its estimated location based on past knowledge. Hence, for the occasional circumstance when there are no matches between frames in the look-ahead buffer and those in the F_{base} database, matching is attempted with previously dropped frames. If one of those frames provides sufficient matching features, it will be treated as a new frame and both will be updated in the filter. This mechanism ensures crucial information can be added back to the filter at any time to mitigate the undesirable situation in USAR when rescuers become rescuees themselves.

B. EIF SLAM

Computational advantages of using an Extended Information Filter rather than an Extended Kalman Filter are now well known, particularly in situations where excessive information is to be processed. This work employs an EIF that maintains all the features as well as the entire sequence of camera poses in the state vector. New camera poses are initialized with respect to the best matching frame at a known pose and measurement updates are additive in the information form. The sensor package is assumed to operate in full 6 DoF without a process model, therefore the formulation of the filter becomes simpler and results in a naturally sparse information matrix.

For full 6 DoF SLAM, the state vector X contains a set of 3D features and a set of camera poses. The camera poses are represented as

$$(x_C, y_C, z_C, \alpha_C, \beta_C, \gamma_C) \tag{1}$$

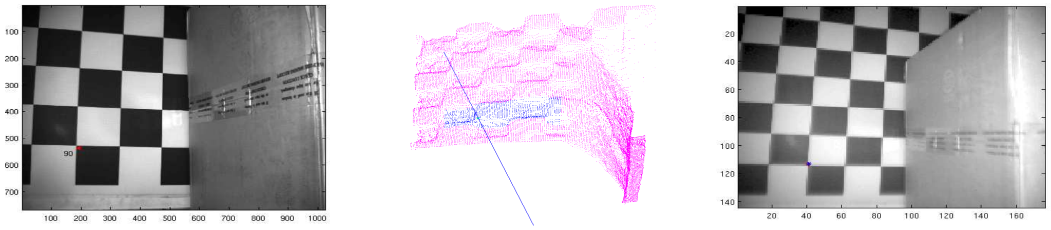
in which α_C, β_C and γ_C represents the ZYX Euler angle rotation and the corresponding rotation matrix is referred to as $RPY(\alpha_C, \beta_C, \gamma_C)$. A 3D point feature in the state vector is represented by

$$(x_F, y_F, z_F) \tag{2}$$

expressed in the global coordinate frame.

Let i represent the information vector and I be the associated information matrix. The relationship between the estimated state vector $\hat{X}$, the corresponding covariance matrix P, the information vector i, and the information matrix I is

$$\hat{X} = I^{-1}i, P = I^{-1} \tag{3}$$



(a)

(b)

(c)

Fig. 2. Example of feature 2D to 3D registration. 2(a) 2D SIFT features are firstly extracted from the traditional camera image before their 3D locations are registered. In the example only feature no. 90 is shown for clarity. 2(b) Bearing to feature intersecting with SwissRanger 3D point cloud. Blue area indicates area of search in the point cloud around the bearing line depending on effective SwissRanger resolution around that range. 2(c) Potential 3D feature projected back onto SwissRanger intensity image for comparison. It can be seen how the feature location does not appear to match exactly that of the pin-hole camera due to significant resolution differences and calibration accuracy, but is is the closest match within the given measurement uncertain boundaries. A subsequent 3D outlier removal step will also filter out potential mismatches.

The first camera pose is chosen as the origin of the global coordinate system. At time $t = 0$, the state vector X contains only the initial camera pose $[0, 0, 0, 0, 0, 0]^T$, and the corresponding 6×6 diagonal information Matrix I is filled with large diagonal values representing the camera starting at a known position.

The observation model provides an estimation of the position of the new features by

$$\begin{pmatrix}\hat{x}_F\\ \hat{y}_F\\ \hat{z}_F\end{pmatrix} = \begin{pmatrix}\hat{x}_C\\ \hat{y}_C\\ \hat{z}_C\end{pmatrix} + \left(RPY(\hat{\alpha}_C, \hat{\beta}_C, \hat{\gamma}_C)^T\right)^{-1} \begin{pmatrix}x_L\\ y_L\\ z_L\end{pmatrix} \tag{4}$$

where x_L, y_L and z_L are the feature location expressed in the local reference frame. In the update step, the information vector and information matrix update can be described by

$$\begin{aligned} I(k+1) &= I(k) + \nabla H_{k+1}^T Q_{k+1}^{-1} \nabla H_{k+1} \\ i(k+1) &= i(k) + \nabla H_{k+1}^T Q_{k+1}^{-1} [z(k+1) - \\ &\quad - H_{k+1}(\hat{X}(k)) + \nabla H_{k+1} \hat{X}(k)] \end{aligned} \tag{5}$$

where Q_{k+1} is the covariance matrix of the observation noise w_{k+1} and $z(k+1)$ is the observation vector. The corresponding state vector estimation $\hat{X}(k+1)$ can be computed by solving a linear equation

$$I(k+1)\hat{X}(k+1) = i(k+1) \tag{6}$$

In error covariance form, the determinant of the $N \times N$ covariance matrix indicates the volume of the N-dimensional uncertainty polyhedron of the filter. The smaller the volume, the more confident the filter is about its estimation. As the information matrix has an inverse relationship with the covariance matrix, as described by (3), the maximally informative frame must update the information matrix to have the largest determinant. We use the natural logarithm of the information matrix determinant, denoted as $log(det(I(k+1)))$, as the measurable quantity of this information update. As described in section IV-A, in a sequence of overlapped images containing common features, each image is evaluated with respect to the base frame database, F_{base}, with same number of new features. Thus, in order to proceed with the actual update of the filter, the pose corresponding to the frame such that $log(det(I(k+1)))$ is maximized becomes the one added to the filter. An empirical threshold to gauge the update quality is also defined based on the desired coarseness of the map. When the maximum determinant is smaller than this threshold, meaning there is little information gain in updating the filter with the current sequence in the look-ahead buffer, all the frames in the sequence are updated to maximize the information gain.

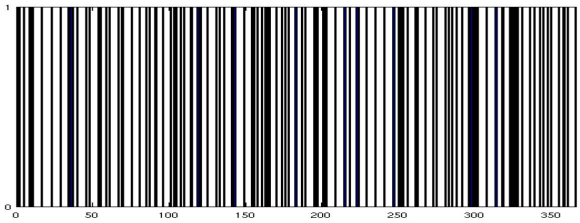


Fig. 3. Distribution of selected frames over the entire data sequence.

V. Results

For experimental evaluation, scans were collected from a mock-up USAR arena measuring 6×3 meters approximately, depicted in Fig. 5(d). The sensor package was hand-held and operated at a combined sampling rate of approximately $5Hz$ in a modern laptop, partly due to the extra time consumed by saving data to the hard drive for later batch-processing. The sensor package is maneuvered in relatively slow motion and it is therefore assumed to produce synchronized data from both cameras at this frequency. 366 scans were collected in about 1 minute and 13 seconds. By applying the proposed approach, 118 out of the 366 scans were automatically selected to be added to the filtering process based on the estimator's uncertainty belief, which represents around 30% of the acquired data. The final distribution of the selected frames over the entire data sequence is depicted in Fig. 3. The state vector ended up containing a total of 2550 elements (118 camera poses (6D) and 614 features (3D)). The final dense 3D map constructed by superimposing the local point clouds to the filtered camera trajectory is shown in Fig. 5(a), where texture has been projected back into the cloud points visible in the the 2D camera images (field of view of the SwissRanger is very close but slightly

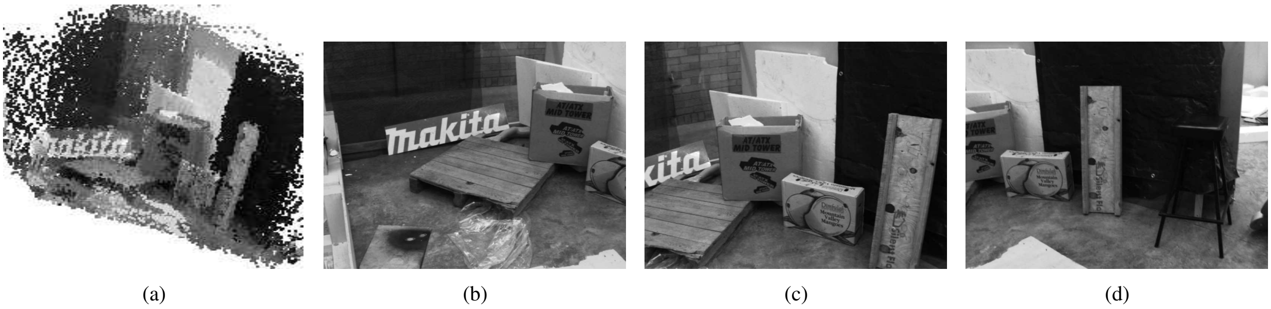


Fig. 4. Partial 3D map reconstructed for the corner area covered by frames 94 to 140 (46 frames). Only 14 frames were actually processed in the filter which was sufficient to produced the highly detailed map seen in 4(a). 4(b) Frame 94. 4(c) Frame 118. 4(d) Frame 140.

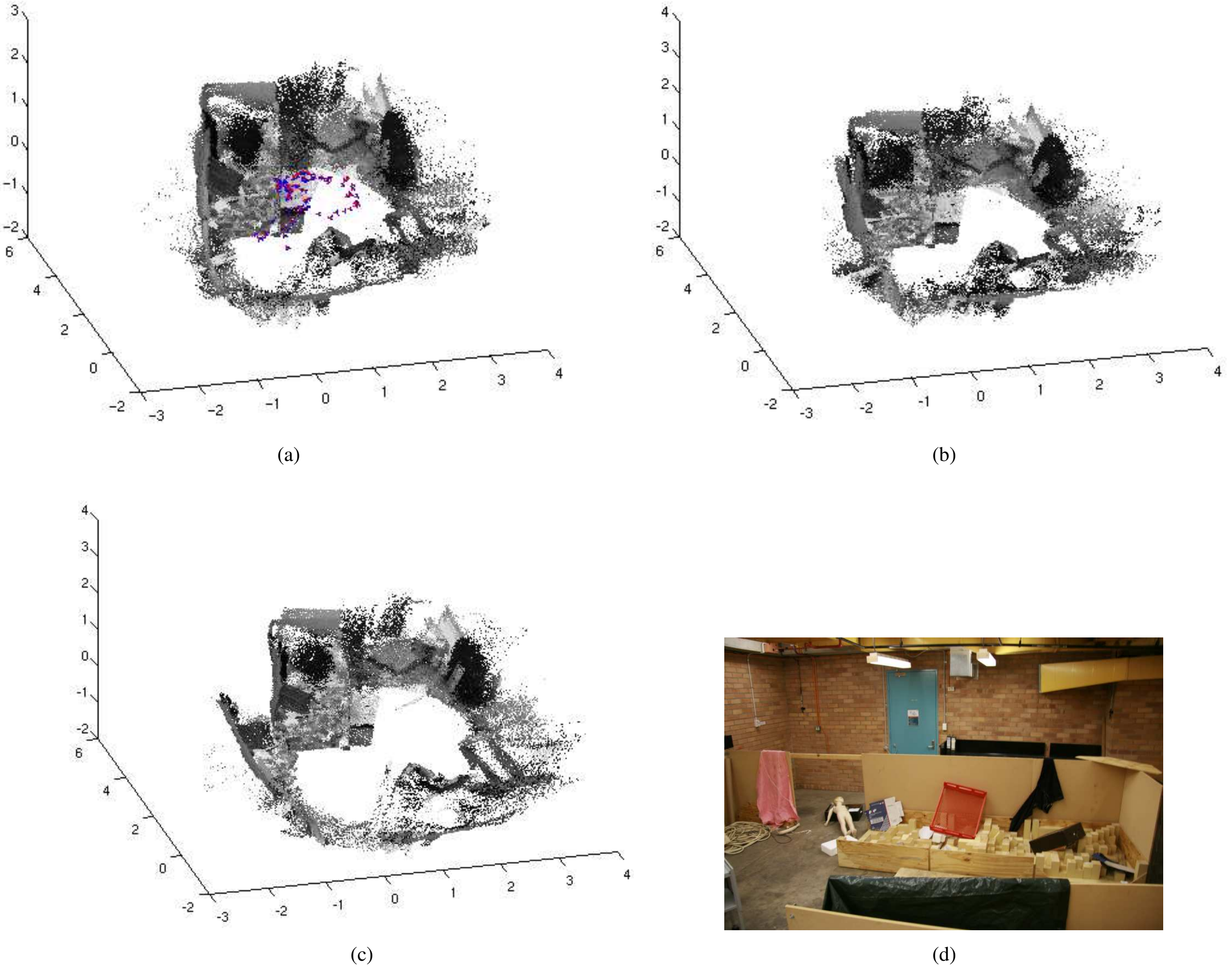


Fig. 5. 3D map obtained from filtering of 118 (out of 366) frames of the 6 × 3 meters search and rescue arena. 5(a) Resulting 3D point cloud map of the entire area superimposed on the selected estimated camera poses, as viewed from one the corners. 5(b) EIF SLAM result obtained by incorporating all frames. 5(c) Final 3D point cloud reconstructed by direct 3D registration between consecutive frames. Overall view of the search and rescue arena is showed in 5(d).

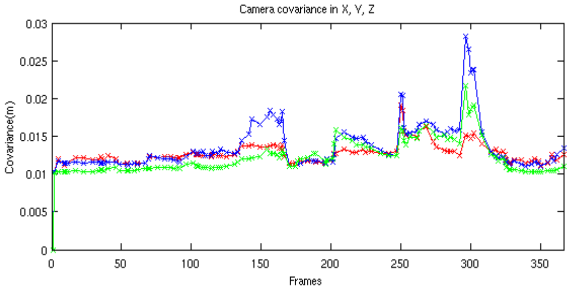


Fig. 6. x_C(red), y_C(blue), z_C(green) covariance of camera poses. Cross marks the frames that are incorporated in the filter.

larger than the field of view of the traditional camera). A more detailed reconstruction of a small corner section and some of its constituent frames is displayed in Fig. 4. Camera covariances x_C, y_C and z_C are further illustrated in Fig. 6 to show to some extent (ground truth is not feasible given the nature of the experiment) the bounded nature of the errors and the filter corrections. Results were also collected for comparison by implementing a standard EIF SLAM and the direct registration of 3D features between consecutive frames over the full data set. The 3D maps constructed by these methods are presented in Fig. 5(b) and 5(c) respectively. Direct registration, as expected, exhibits a large accumulated error of more than 2 meters in the X direction. While full SLAM, despite the extended computing time it consumed, provides a comparably decent output.

From all the experiments presented, we can conclude that the proposed algorithm is most valid for handling real data sets and constitutes a significant step in visually improving the map quality of 3D unstructured environments in an efficient manner.

VI. Conclusion and Future Work

We have presented an approach for producing 3D texture-rich map with a combination of vision and range sensors. The proposed algorithm not only produces consistent SLAM outputs but also dynamically incorporates observations into the estimation process for efficient 3D navigation in unstructured terrain.

Unlike most conventional fixed time step or fixed displacement approach, our proposed technique exhibits the ability to fuse the minimal information required based on the robot uncertainty belief and the perceived quality of the observation. Results have shown that by gauging the information gain in each frame, we can automatically incorporate the most apt observations for the purpose of SLAM and extract comprehensive findings about the collective environment we intend to explore.

The proposed knowledge-driven algorithm can be regarded as an apparent trade off between computational efficiency and information loss. We believe it is critical for intelligent systems in the field to distinguish what constitutes relevant information from what is not of the same significance within the realm of the objective at hand. It is not hard to imagine how, for a USAR robot exploring inside a collapsed building, it is of higher priority to place exit points, stairways, windows, large cavities etc. with relative accuracy, rather than yielding undue emphasis on generating perfectly straight walls. Providing the capacity to attribute a measure of relevance to the information attained for a given objective seems to us like a noteworthy step.

Although the current results appear promising, there are still some limitations to the proposed strategy due to factors such as the limited robustness of SIFT to more dramatic changes in view angles, or the noisy measurements returned by the SwissRanger when it encounters some types of surfaces (such as glass), as well as a limited non-ambiguous range of operation. Also, while the proposed algorithm has been much improved in terms of computation efficiency compared to earlier efforts, exploration of much larger areas is still a challenge to be met.

Acknowledgment

This work is supported by the Australian Research Council (ARC) through its Centre of Excellence programme, and by the New South Wales State Government.

References

[1] M. Fischler, R. Bolles, "RANdom SAmpling Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography", *Communications of the Association for Computing Machinery*, vol. 24, pp. 381–395, 1981.

[2] K. S. Arun, T. S. Huang, S. D. Blostein, "Least Square Fitting of Two 3-D Point Sets", *IEEE Pattern Analysis and Machine Intelligence*, vol. 9(5), pp. 698–700, 1987.

[3] F. Lu, E. Milios, "Globally Consistent Range Scan Alignment for Environment Mapping", *in Autonomous Robots*, vol. 4(4), pp. 333–349, 1997.

[4] 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

[5] T. Oggier, M. Lehmann, R. Kaufmann, M. Schweizer, M. Richter, P. Metzler, G. Lang, F. Lustenberger, N. Blanc, " An All-Solid-State Optical Range Camera for 3D Real-Time Imaging with Sub-Centimeter Depth Resolution (SwissRanger)", *Proceedings of the SPIE*, vol. 5249, pp. 534–545, 2004.

[6] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints" *Int. Journal of Computer Vision*, vol. 60(2), pp. 91–110, 2004.

[7] R. M. Eustice, O. Pizarro, H. Singh, "Visually Augmented Navigation in an Unstructured Environment Using a Delayed State History", *Proceedings of the IEEE Int. Conf. on Robotics and Automation*, pp. 25–32, 2004.

[8] R. M. Eustice, H. Singh, J. J. Leonard, "Exactly Sparse Delayed-Sate Filters", *Proceedings of the IEEE Int. Conference on Robotics and Automation*, pp. 2417–2424, 2005.

[9] K. Ohno, S. Tadokoro, "Dense 3D Map Building Based on LRF Data and Color Image Fusion", *Proceedings of the IEEE Int. Conf. on Intelligent Robot and Systems*, pp. 1774–1779, 2005.

[10] P. Newman, K. Ho, "SLAM - Loop Closing with Visually Salient Features", *Proceedings of the IEEE Int. Conf. on Robotics and Automation*, pp. 644–651, 2005.

[11] D. M. Cole, P. M. Newman, "Using Laser Range Data for 3D SLAM in Outdoor Environments", *Proceedings of the IEEE Int. Conf. on Robotics and Automation*, pp. 1556–1563, 2006.

[12] L. P. Ellekilde, S. Huang, J. Valls Miró, G. Dissanayake, "Dense 3D Map Construction for Indoor Search and Rescue", *Journal of Field Robotics*, vol. 24(1/2), pp. 71–89, 2007.

[13] W. Zhou, J. Valls Miró, G. Dissanayake, "Information Efficient 3D Visual SLAM in Unstructured Domains", *Proceedings of the IEEE Int. Conf. on Intel. Sensors, Sensor Networks and Information Processing*, pp. 323–328, 2007.