SLAM using Visual Scan-Matching with Distinguishable 3D Points

Federico Bertolli, Patric Jensfelt, and Henrik I. Christensen
Centre for Autonomous Systems
Royal Institute of Technology, Stockholm, Sweden
fedebsw@yahoo.it, [patric,hic]@nada.kth.se

***Abstract*— Scan-matching based on data from a laser scanner is frequently used for mapping and localization. This paper presents an scan-matching approach based instead on visual information from a stereo system. The Scale Invariant Feature Transform (SIFT) is used together with epipolar constraints to get high matching precision between the stereo images. Calculating the 3D position of the corresponding points in the world results in a visual scan where each point has a descriptor attached to it. These descriptors can be used when matching scans acquired from different positions.**

Just like in the work with laser based scan matching a map can be defined as a set of reference scans and their corresponding acquisition point. In essence this reduces each visual scan that can consist of hundreds of points to a single entity for which only the corresponding robot pose has to be estimated in the map. This reduces the overall complexity of the map.

The SIFT descriptor attached to each of the points in the reference allows for robust matching and detection of loop closing situations. The paper presents real-world experimental results from an indoor office environment.

I. Introduction

Simultaneous Localization and Mapping, or SLAM, is the process of concurrently building a map of the environment and using that map for estimating the position of the robot. This is a key component in an autonomous mobile robot system and as such has attracted a lot of attention in the robotics community.

SLAM based on range sensors in indoor, structured, environments is now considered to be a mature technology within the research community. As an example, several systems using ultra sonics sensors [1]–[3] and lately more often with a laser scanner [4]–[9] have been presented. Range sensors provide mostly a geometric interpretation of the environment. Furthermore, the laser scanner is comparably expensive which makes it unfit for many applications, especially high volume type applications.

In the last few years the focused has shifted from using laser scanners to visual systems [10]–[19]. The amount of information available in images far surpasses that what a laser scan provides but much more involved algorithms are needed to extract that information. However, a camera system will probably offer a much more cost efficient solution as computational power gets cheaper, and may therefore be applicable even to consumer type products.

In the computer vision community the problem of structure from motion (SFM) has been studied for quite some time and even before that vision based reconstruction was studied in the photogrammetry community. The formulation of the SFM problem is similar to SLAM but in SFM the position of the camera is typically estimated using only the visual information whereas in SLAM additional information from odometry often is incorporated. Where SFM typically is performed offline by batch processing, SLAM aims at running in real time on a robot with all the challenges that this brings in terms of computational complexity, scalability, etc.

Davison was one of the first to address the problem of visual SLAM. In [20] a system is presented that uses a stereo rig to detect and fixate on visual landmarks. The extended Kalman filter provides the framework for fusing odometry and visual information as well as information from accelerometers. The robot is able to build a map while traveling over uneven terrain. In more recent work Davison has focused on single camera SLAM and in particular on how SLAM can be performed without any information from odometry. This is important when the camera is hand held or mounted on a human for example [21].

Se, Lowe and Little [13], [14] use the *Scale Invariant Feature Transform* or SIFT invented by Lowe [22] in an EKF-based implementation of SLAM. In [23] a Rao-Blackwellised particle filter is instead applied to SLAM using SIFT features.

Karlsson et al. [16], [17] use a monocular camera and combine this with SIFT features in the so called vSLAM algorithm. The 3D position of points is estimated through *Structure from Motion* using three consecutive frames. The collection of 3D points along with their SIFT descriptors defines a landmark in their map. The mapping process creates new landmarks when there is no correspondence between the SIFT features in the current image and the previous landmarks.

Visual SLAM has also been used in under-water applications. In [24] a downward looking camera that overlooks the sea floor is used. The system uses an augmented state Kalman filter to estimate the position of the vehicle at the current position and the past trajectory. Each position in the old trajectory is represented by one image. The overlap between different images can be used to create measurements.

In [25] a system for so called visual odometry is presented. This term refers to estimating the motion of a camera system based only on visual information. In [26] stereo data in combination with ICP is used to estimate the 6 DoF motion of a robot that moves in rocky terrain. No map is made in either

of these cases.

In this paper we work with a stereo camera and use the SIFT descriptor to match interest points between the two camera images. What sets this work apart from for example [13] that also use SIFT features is that we do not treat the individual points as landmark but instead treat all the points that are matched between the two stereo images as one 3D scan much like a laser scan. The whole cluster of SIFT points can in this way be used to identify the scan when matching new scans to old.

The idea in this paper is to use the visual scans like in the abundance of work with laser based scan matching [4], [6], [27]. That is, the raw scans together with the positions from where they were acquired define the map. Furthermore, just like in [24] we will use the EKF framework with an augmented state representation to demonstrate our idea. However, any of the algorithms used for laser based scan matching can be applied here as well. In contrast to the laser based counter part our scans have very strong discriminative power thanks to the SIFT points. The rest of the paper presents one possible implementation of visual scan-matching to illustrate the concept.

II. Visual Scan

As mentioned in the Introduction we use the *Scale Invariant Features Transform* or SIFT points by David Lowe [28]. These has become very popular both in the vision community for example for object recognition and in robotics for example as landmarks for SLAM. The strength of the SIFT is that it is invariant to scale changes and image rotation up to approximately 30° which allows for robust matching even when the view point has changed.

This section just describes one way to extract and define the visual scans. The standard implementation available online from David Lowe is used for extracting the SIFT key points. Points are detected as minimas and maximas of differences of Gaussians across different images scales. Along with the location in the images, each point has a feature vector associated with it. This feature vector consists of the image pixel gradients calculate in 8 direction in a 4x4 grid around the point, which results in 4x4x8=128 values. In addition the scale and dominant orientation for each point is stored.

In particular the descriptor of 128 elements and the orientation associated with each feature, are very robust to image changes and allows to track and recognize a point in a sequence of consequent images, acquired along the robot path. Typically the Euclidean distance between the two descriptors are used to when matching two SIFT points.

A. Creating a Visual Scan

To create the visual scans we have to match interest points between the two camera images acquired by the stereo system. To make the matching between the left and right camera images easier we make sure that the images are rectified. We exploits this and the fact that the geometry of stereo system (baseline of 160mm) and the intrinsic parameters of the camera, are known in the matching. An alternative would be to use motion stereo to create the visual scans.

The Euclidean distance between the descriptors and the difference between the orientations is used to verify the matching. The threshold for matching SIFT points p and p' is set to 8% between the descriptors δ, δ' and 20° between the orientations θ and θ'. For each point we pick as the matching point the one closest to the current. Formally, the matching can be expressed as:

1) $\mid \theta - \theta' \mid < 20$
2) $\parallel \delta - \delta' \parallel^2 < \parallel \delta - \delta'' \parallel^2, \forall \delta'' \in$ SIFT on the same row
3) $\parallel \delta - \delta' \parallel^2 < 62000$

where $62000 \simeq 22^2 \cdot 128$ means a tolerance of about 8% for each couple of values of the two descriptors.

We perform matching both left to right and right to left and keep only those points that has a one-to-one matching.

For each of the matching SIFT points we then calculate the 3D position. These 3D position along with the SIFT descriptor from, in our case, the left image defines our visual scan.

B. Matching Visual Scans

For each new image pair we produce a new visual scan and match this to existing scans. This matching is performed in steps. First we find all existing scans that are close to the current pose of the robot. That is, we select all visual scan which the current scan is likely to be able to match. Then, we use the estimated difference between the robot pose corresponding to the existing scan and the current pose to predict the image coordinates for the points from the existing scan in the current image.

Currently, there is a fixed search window around the predicted position of each point. The dimensions of these windows are 40×14 pixels which allows for an uncertainty of $\pm 8^\circ$ (horizontally) and $\pm 3^\circ$ (vertically). For the matching between features in consequent images, similar conditions as in the stereo matching are used:

1) $\parallel \delta_1 - \delta_2 \parallel^2 < \parallel \delta_1 - \bar{\delta}_2 \parallel^2, \forall \bar{\delta}_2 \in$ SIFT in the tolerance window
2) $\parallel \delta_1 - \delta_2 \parallel^2 < 74000$

The condition on orientation is not used here and a bit more margin is used when matching the descriptors, $74000 \simeq 24^2 \cdot 128$ means a tolerance of about 10%.

Matching is performed between the points in the two scans using the descriptors from the left images. Also here a 1-1 match is required, that is, the matching is performed in both directions and only those matches that are consistent are considered correct.

III. Estimation of Relative Displacement

Given a set of matches between the current scan and an existing scan the relative displacement between the scans must be computed. This serves as the measurement in the system.

We we use the method by Kanatani [29] to solve for the absolute orientation and to get the pose between two sets of

points, represented in two reference systems:

$$\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N} \|\mathbf{w}_i - (\mathbf{R}\mathbf{w}'_i + \mathbf{t})\|^2 \quad (1)$$

If $\bar{\mathbf{c}}$ and $\bar{\mathbf{c}}'$ are the *centroids* of the two sets and $\bar{\mathbf{w}}_i = \mathbf{w}_i - \bar{\mathbf{c}}$, $\bar{\mathbf{w}}'_i = \mathbf{w}'_i - \bar{\mathbf{c}}'$ are the two point sets translated with the corresponding centroid, it is possible to solve an equation equivalent to Eq. 1 without $\mathbf{t}$:

$$\min_{\mathbf{R}} \sum_{i=1}^{N} \|\bar{\mathbf{w}}_i - \mathbf{R}\bar{\mathbf{w}}'_i\|^2, \quad (2)$$

and finally to solve for the translation $\mathbf{t}$ with:

$$\mathbf{t} = \bar{\mathbf{c}} - \mathbf{R}\bar{\mathbf{c}}'. \quad (3)$$

Since the data is corrupted by noise like inaccurate feature localization, intrinsic error in triangulation and false correspondences, it's not possible to use the minimization in Eq. 2 without any system to compensate these errors. To address this problem we use the RANSAC algorithm (Random Sample Consensus) [30].

RANSAC is a voting protocol that permits good results with up 50% of outliers. This algorithm works with a set of elements S and a model characterized by ξ parameters. S is related to the couples of correspondent points and ξ to the pose $\mathbf{R}$, $\mathbf{t}$. RANSAC requires six parameters: P, p, ε, N, T, t. P is the probability to randomly get p inliers in N tries ($P = 0.999$) and p is the minimum number of points, to compute $\mathbf{R}$ and $\mathbf{t}$ with Kanatani knowing the intrinsic parameters of the camera ($p = 5$). ε is the unknown percentage of outliers (experimentally found $\varepsilon = 0.35$). The parameter T is the unknown absolute number of inliers (in percentage $T = 1 - \varepsilon$) and t is the threshold for the vote.

Now it is possible to find N with:

$$\begin{aligned} P &= 1 - (1 - (1-\varepsilon)^p)^N \\ 0.999 &= 1 - (1 - (1 - 0.35)^5)^N \\ \Rightarrow N &\simeq 56 \end{aligned}$$

The value of t depends on the average distance of each couple of points (z', z'') to the camera. The following equation gives the uncertainty of a general triangulated point:

$$\Delta z = \frac{\sqrt{2} \cdot z^2 \cdot \Delta v}{b \cdot f} \cdot dimPixelCCD \quad (4)$$

where $z = \frac{z'+z''}{2}$, b is the baseline, f is the focal length and δv is the precision in the point coordinates. This is correct if the disparity of two corresponding points is perfect. To account for more realistic cases an experimental constant $unZ = 1.88$ is multiplied to Δz in order to increase the uncertainty. So, finally $t = unZ \cdot \Delta z$. The relative displacement is now found with 56 iterations of RANSAC and Kanatani. The pose that accumulates the most votes is chosen and the corresponding matches are used to calculate the final estimate of the relative displacement.

To remove highly uncertain matches we use a threshold of $T = 65\%$ and also require that we have at least 10 matching points.

A. Visual Reference Scans

In order to implement visual scan-matching we need to store visual reference scan and thus also define which scans to store. The decision on what to turn into a reference scan is delayed one step. That is, at step k we decide if the scan acquired in step $k-1$ should become a reference scan. The scan in step $k-1$ is turned into a reference scan if in step k we are unable to match to any existing reference scan but can match to the previous scan. Notice here that we, by the definition of the matching rules, cannot match to a scan that has too few points. Therefore we will not turn scans with too few points into reference scans.

IV. Map Estimation

In this paper we use an augmented state Kalman filter for estimating the current robot position and the position of the visual reference scans, similar to [24].

The state vector starts out containing only the robot pose,

$$\mathbf{x}(k) = \left(\mathbf{x}_r(k)\right). \quad (5)$$

To allow for delaying the decision about turning a visual scan into a visual reference scan one step we also keep the previous robot pose in the state vector, i.e. disregarding the reference scans the state vector contains,

$$\mathbf{x}(k) = \begin{pmatrix} \mathbf{x}_r(k) \\ \mathbf{x}_r(k-1) \end{pmatrix}. \quad (6)$$

When the previous scan is flagged as a reference scan we simply let the second state in the state vector, the previous pose and also the current estimate of the pose of the new reference scan, transition into a reference position, i.e.

$$\mathbf{x}(k) = \begin{pmatrix} \mathbf{x}_r(k) \\ \mathbf{x}_r(k-1) \end{pmatrix} \Rightarrow \mathbf{x}(k+1) = \begin{pmatrix} \mathbf{x}_r(k+1) \\ \mathbf{x}_r(k) \\ \mathbf{x}_1(k+1) = \mathbf{x}_r(k-1) \end{pmatrix}, \quad (7)$$

where $\mathbf{x}_r(k+1)$ is the new robot pose after the next step has been taken.

This process continues by augmenting the state vector for every new reference scan,

$$\mathbf{x}(k) = \begin{pmatrix} \mathbf{x}_r(k) \\ \mathbf{x}_r(k-1) \\ \mathbf{x}_N(k) \\ \vdots \\ \mathbf{x}_2(k) \\ \mathbf{x}_1(k) \end{pmatrix}, \quad (8)$$

where N is the number of reference scans. When revisiting areas, new visual reference scans do not need to be added if the robot can navigate with respect to already existing scans.

This augmented state Kalman filter implementation is just one of many possible ways to realize visual scan-matching. Even though the representation with reference scans reduces the number of states in the state vector with respect to keeping all the individual SIFT points as in [14] the Kalman filter still scales badly. As an example the FastSLAM algorithm could be used as in [6] with laser based scan matching.

Fig. 1. Left: The experimental platform, the PowerBot Dumbo from ActivMedia. Right: A close up of the Videre stereo rig used in the experiments.

V. Experimental Evaluation

We performed the experimental evaluation on a PowerBot robot equipped with a Videre STH−MDSC2−VAR−C stereo rig. The robot and a close up of the stereo system can be seen in Fig. 1.

For the evaluation we drove the robot around in our indoor office environment along a trajectory starting in one room, passing through a corridor section to another room and then back ending up at almost exactly the same position as the start position. Figure 2 shows some images from the the environment.

Fig. 2. Some images from the environment. The top row shows to different views from the room where the robot start. On the lower row the first image shows the view when driving out of the starting room and the second shows the rather sparse corridor.

Figure 3 shows snapshots of the positions of the reference scan poses along the trajectory. The uncertainty in the reference positions are illustrated with uncertainty ellipses. Notice how drift causes the estimate to deteriorate but how at the end when the robot re-observes some of the initial reference scans this is corrected.

In Figure 4 a histogram is shown over the number of points per scan. The visual scans contains up to 206 scan points in this experiment. In total there are 126 reference scans and a total of 8333 points in the map. Using the reference scans has thus reduced the number of landmarks from 8333 to 126, i.e., a factor of 66. This is not entirely true though as some of the points will be represented in more than one reference scan but it is clearly a very large reduction and at the same time the map still contains all the descriptive power of using all points.

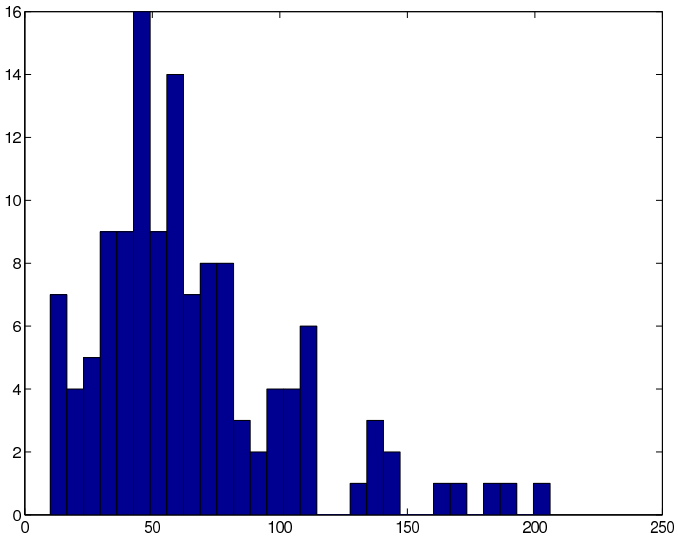
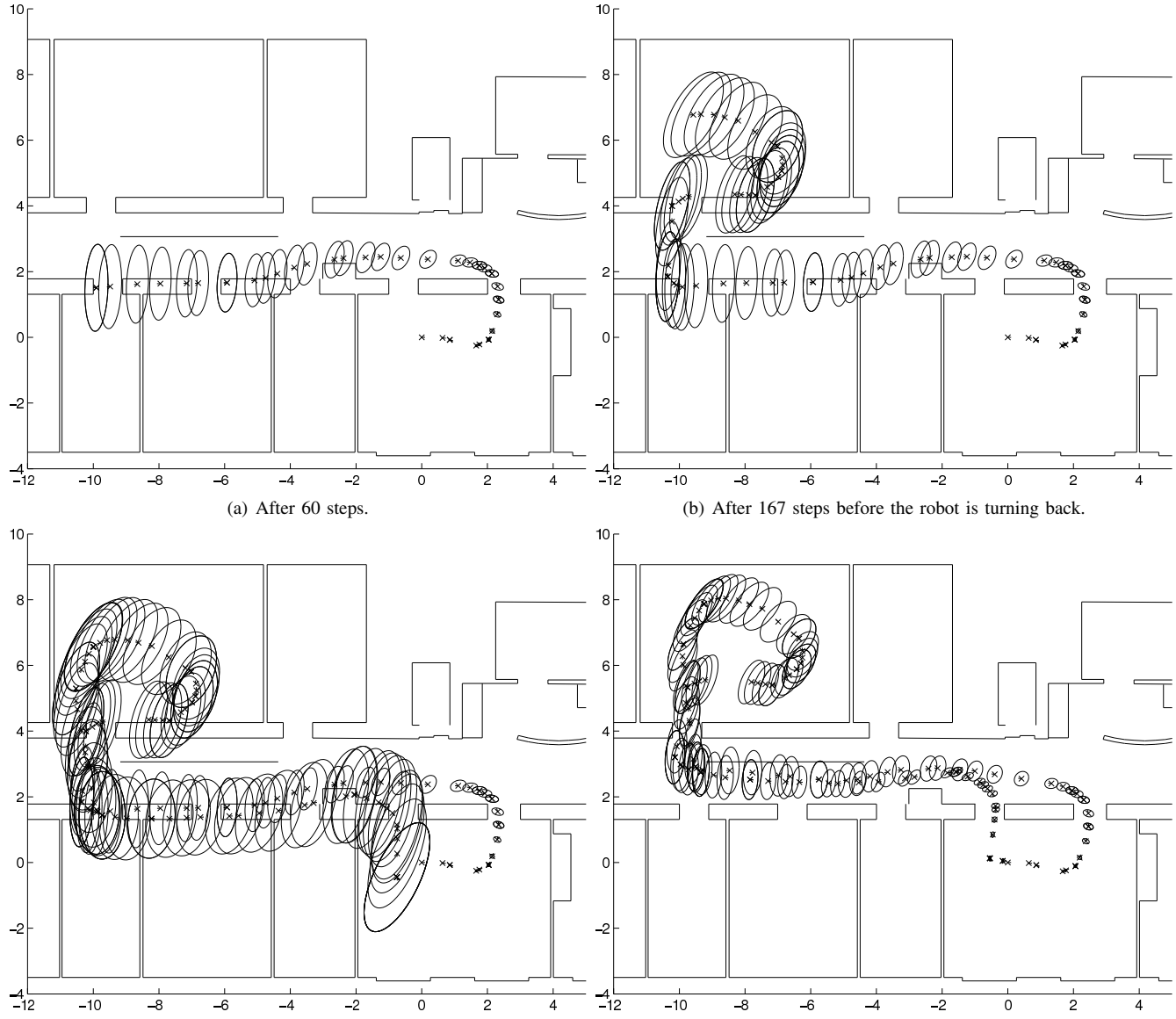


Fig. 4. Histogram showing the number of points per scan.

Figure 5 shows the final map with all points from the different reference scans overlayed. Notice how the scans are nicely aligned.



(a) After 60 steps.

(b) After 167 steps before the robot is turning back.

(c) After 266 just before observing some of the first reference scans again.

(d) The final estimate, after 272 steps.

Fig. 3. The figure shows the position and uncertainty of the reference scan positions at different time during the building of the map. Notice how the map is corrected when the loop is closed and the robot successfully matches to some of the initial reference scans.

VI. Summary and Conclusion

In this paper we have introduced the concept of visual scan matching. The idea is to build upon the success that laser based scan matching has had. This way we inherit the strong advantages of scan matching such as the representation flexibility (the sensor data itself is the representation) and combine that with the advantages of using vision which adds the ability to add appearance to the data association process.

Some initial experimental results were presented where a stereo rig was used to estimate the 3D position of the points detected and described by the Scale Invariant Feature Transform or SIFT.

Acknowledgment

This work was partially sponsored by the SSF through its Centre for Autonomous Systems (CAS) and the EU as part of the project CoSy IST-2004-004450. The support is gratefully acknowledged.

References

[1] S. Thrun, W. Burgard, and D. Fox, "A probabilistic approach to concurrent mapping and localization for mobile robots." in *Machine Learning and Autonomous Robots (joint issue), 31(5):1-25*, 1998.

[2] K. S. Chong and L. Kleeman, "Feature-based mapping in real, large scale environments using an ultrasonic array," in *Proc. of FSR*, 1997, pp. 538–545.

[3] J. Tardós, J. Neira, P. Newman, and J. Leonard, "Robust mapping and localization in indoor environments using sonar data," *IJRR*, vol. 21, no. 4, pp. 311–330, Apr. 2002.

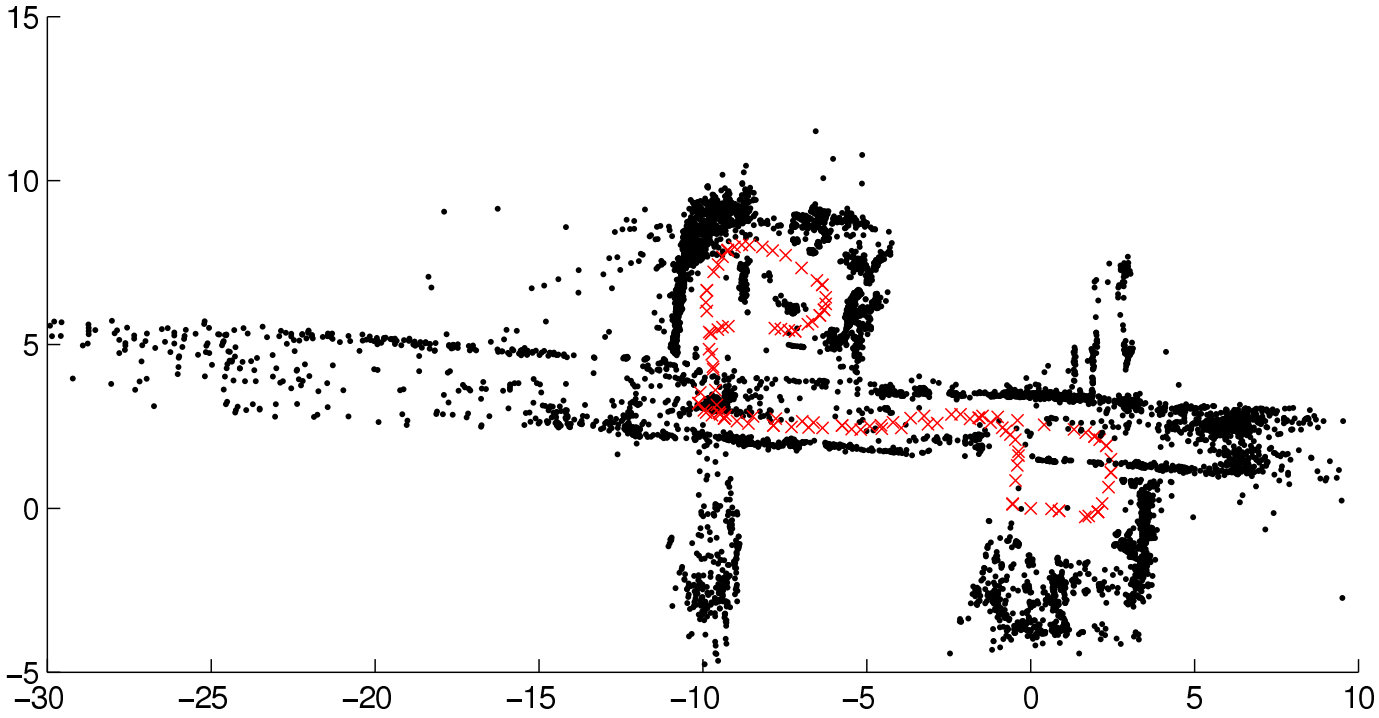


Fig. 5. The resulting map with with all the reference scans overlayed. The map consists of 8333 points in total but only 126 visual reference scan. The 'x' mark the position of the robot pose corresponding to the reference scans.

[4] J. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments." in *Proc. of the CIRA*, 1999, pp. 318–325.

[5] S. Thrun, W. Burgard, and D. Fox, "A real−time algorithm for mobile robot mapping with applications to multi−robot and 3d mapping." in *Proc. of the ICRA*, 2000, pp. 321–328.

[6] D. Hähnel, W. Burgard, D. Fox, and S. Thrun, "An efficient FastSLAM algorithm for generating maps of large-scale cyclic environments from raw laser range measurements," in *Proc. of the IROS*, vol. 1, Oct. 2003, pp. 206–211.

[7] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem," in *Proc. of the AAAI*, Edmonton, Canada, 2002.

[8] S. Thrun, Y. Liu, D. Koller, A. Ng, Z. Ghahramani, and H. Durrant-Whyte, "Simultaneous localization and mapping with sparse extended information filters," *IJRR*, 2004.

[9] M. Bosse, P. Newman, J. Leonard, and S. Teller, "SLAM in large-scale cyclic environments using the Atlas framework," *IJRR*, vol. 23, no. 12, pp. 1113–1139, 2004.

[10] C. Harris, *Geometry from visual motion.*, A. Blake and A. Yuille, Eds. Active Vision, MIT Press, 1992.

[11] J. J. Little, J. Lu, and D. Murray, "Selecting stable image features for robot localization using stereo," *Proc. of the IROS*, pp. 1072–1077, 1998.

[12] J. Mallon, O. Ghita, and P. Whelan, "Robust 3d landmark tracking using trinocular vision," in *OPTO-Ireland: SPIE's Regional Meeting on Optoelectronics, Photonics and Imaging, Galway*, 2002.

[13] S. Se, D. Lowe, and J. Little, "Vision−based mobile robot localization and mapping using scale−invariant features," in *Proc. of the ICRA*, 2001, pp. 2051–58.

[14] ——, "Mobile robot localization and mapping with uncertainty using scale−invariant visual landmarks," vol. 21, no. 8, 2002.

[15] M. A. Garcia and A. Solanas, "3d simultaneous localization and modeling from stereo vision," in *Proc. of the ICRA*, 2004.

[16] N. Karlsson, E. D. Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M. E. Munich, "The VSLAM algorithm for robust localization and mapping," in *Proc. of the ICRA*, 2005, pp. 24–29.

[17] L. Goncalves, E. D. Bernardo, D. Benson, M. Svedman, J. Ostrowski, N. Karlsson, and P. Pirjanian, "A visual front−end for simultaneous localization and mapping," in *Proc. of the ICRA*, 2005.

[18] J. Folkesson, P. Jensfelt, and H. Christensen, "Vision slam in the measurement subspace," in *Proc. of the ICRA*, Apr. 2005.

[19] P. Jensfelt, D. Kragic, J. Folkesson, and M. Björkman, "A framework for vision based bearing only 3D SLAM," in *Proc. of the ICRA*, Orlando, FL, May 2006.

[20] A. J. Davison and N. Kita, "3D simultaneous localisation and map-building using active vision for a robot moving on undulating terrain," in *Proc. of the CVPR*, 2001.

[21] A. J. Davison, W. Mayol, and D. W. Murray, "Real-time localisation and mapping with wearable active vision," in *Proc. of the ISMAR*, 2003.

[22] D. G. Lowe., "Object recognition from local scale−invariant features," in *Proc. of the ICCV*, 1999, pp. 1150–1157.

[23] R. Sim, P. Elinas, M. Griffin, and J. J. Little, "Vision-based slam using the rao-blackwellised particle filter," in *Proc. IJCAI Workshop on Reasoning with Uncertainty in Robotics, Edinburgh, Scotland*, July 2005.

[24] R. Eustice, O. Pizarro, and H. Singh, "Visually augmented navigation in an unstructured environment using a delayed state history," in *Proc. of the ICRA)*, vol. 1, Apr., 2005, pp. 25–32.

[25] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proc. of the CVPR*, vol. 1, 2004, pp. 652–659.

[26] A. Milella and R. Siegwart, "Stereo-based ego-motion estimation using pixel tracking and iterative closest point," in *Proc. of the ICVS*, 2006.

[27] F. Lu and E. Milios, "Robot pose estimation in unknown environments by matching 2d range scans," in *Proc. of the CVPR*, June 1994, pp. 935–938.

[28] D. G. Lowe, "Distinctive image features from scale−invariant keypoints," *IJCV*, pp. 91–110, 2004.

[29] K. Kanatani, *Geometric computation for machine vision*. Oxford University Press, Inc., 1993.

[30] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. of the ACM*, vol. 24, pp. 381–395, 1981.