Vision Aided 3D Laser Scanner Based Registration

Henrik Andreasson Achim Lilienthal
Centre of Applied Autonomous Sensor Systems, Dept. of Technology, Örebro University, Sweden

***Abstract*— This paper describes a vision and 3D laser based registration approach which utilizes visual features to identify correspondences. Visual features are obtained from the images of a standard color camera and the depth of these features is determined by interpolating between the scanning points of a 3D laser range scanner, taking into consideration the visual information in the neighbourhood of the respective visual feature. The 3D laser scanner is also used to determine a position covariance estimate of the visual feature. To exploit these covariance estimates, an ICP algorithm based on the Mahalanobis distance is applied. Initial experimental results are presented in a real world indoor laboratory environment.**

***Index Terms*— Registration, Vision**

I. Introduction

Registration or scan-matching is a popular approach to obtain robot relative pose estimates, it is also a very core part of most Simultaneous Localization and Mapping (SLAM) algorithms. Most work that have been published in the past consider 2D-motion in an indoor environment, however, nowadays more attention is directed towards complete 6DOF methods.

Since vision is particularly suited to solve the correspondence problem (data association), vision-based systems have been applied as an extension to laser scanning based SLAM approaches for detecting loop closing. The principle has for example been applied to SLAM systems based on a 2D laser scanner [8] and a 3D laser scanner [12]. When using methods which rely on a weaker correspondence, i.e. point to point distance like in standard ICP [4, 5], a good initial estimate is very important to the robustness of the system. By instead using the strong correspondences visual features can provide, a good initial estimate is not necessary [12].

In addition, the cost of adding a camera is comparably small compared to the cost of a 3D laser scanner. It is further known that vision-based approaches can work in highly cluttered environments where pure laser range scanner based methods fail [14].

This paper presents a registration method which relies on a standard color camera together with a 2D laser scanner mounted on a pan / tilt unit. The method utilizes the strong visual correspondences obtained from the camera which are incorporated with depth information from the 3D laser scanner.

The reminder of this paper is organized as follows. First related work is presented in Section II followed by a description of the suggested method Section III and experimental setup (Section IV). Section V shows preliminary results and finally conclusions and future work are discussed (Section VI).

II. Related Work

The most similar approach to the registraion method suggested in thie paper is the visual loop closing approach by Newman et. al. [12]. The main difference is that we use solely the visual features for registration and not as an initial estimate for a laser based registration method.

Since this work incorporates both visual and 3D laser information there is some overlap in the proposed method compared to approaches using only vision or 3D laser scanner. For example, Lowe's scale invariant feature transform (SIFT) has been used widely in 'pure' vision based solutions [16, 3]. ICP is commonly utilized in 3D laser based registration [13, 17]. Regitration using visual features directly together with an estimate of the corresponding position and position covariance has not been addressed so far to our knowledge.

III. Method

The suggested approach is based on extracting visual features from the image and then using the laser scanner to obtain a position estimate and a position covariance of the visual features. In our current implementation we use the popular SIFT features developed by Lowe [9]. The position covariance for visual features is obtained by using the surrounding laser based range values. I.e. if the detected feature is located on a poster (planar surface), the feature position covariance will be smaller compared to a feature extracted from a branch, for example.

As stated in the previous section, most current approaches to scan registration depends on reasonably accurate initial pose estimates. In our case, the correspondences are solely determined from the visual features and not by their spatial distance only. As a result, no initial pose estimate is required. This makes the method suitable for conditions in which initial pose estimates are not available.

Shortly the registration procedure can be described as follows: first, SIFT features are computed in the planar images recorded with the current scan data $\mathcal{S}_c$ and compared to the SIFT features found in the images belonging to previous scan $\mathcal{S}_p$. Next, the depth values are estimated for all matching feature pairs in $\mathcal{S}_p$ and $\mathcal{S}_c$, using the closest projected 3D laser point as described in Section III-B. Pairs of 3D points corresponding to matching features are then used together with the feature position covariance to obtain the relative pose estimate (see Section III-E).

In a related approach Newman et. al. [12] used SIFT features to detect loop closure events in a 3D SLAM approach. In contrast to their method where SIFT features are used to obtain an initial pose estimate (by determining the essential matrix between two images) and the full point cloud is considered afterwards, registration in our approach is carried out using only 3D points that are associated with matching visual features. By restricting scan matching to 3D points

that were found to correspond by their visual appearance, we believe that the robustness against changes in the environment is improved and more accurate registration can be obtained. We provide evidence for this statement and are currently validating this belief in a thorough ground truth evaluation using a large set of 3D scans.

A. Detecting Visual Correspondences

Given two images I_a and I_b, we extract local visual features using the SIFT algorithm [9] resulting in two sets of features F_a and F_b, corresponding to the two images. Each feature $f_i = \{[X,Y]_i, H_i\}$ in a feature set $F = \{f_i\}$ comprises the position $[X,Y]_i$ in pixel coordinates and a histogram H_i containing the SIFT descriptor.

The feature matching algorithm calculates the Euclidean distance between each feature in image I_a and all the features in image I_b. A potential match is found if the smallest distance is less than 60% of the second smallest distance. This criterion was found empirically and was also used in [7], for example. It reduces the risk of falsely declaring correspondence between SIFT features by excluding cases where a false correspondence is caused by the existence of several almost equally well matching alternatives. In addition, no feature is allowed to be matched against more than one other feature. If more than one candidate for matching is found, the feature with the highest similarity among the candidate matches is selected.

The feature matching step results in a set of feature pairs $P_{a,b}$, with a total number $M_{a,b} = |P_{a,b}|$ of matched pairs.

B. Obtaining the Visual Feature Depth

To obtain the depth estimate r_i^* for SIFT feature f_i the Nearest Range Reading (NR) method [1] is applied.

Image data consist of a set of image pixels $\mathcal{P}_j = (X_j, Y_j, C_j)$, where X_j, Y_j are the pixel coordinates and $C_j = (C_j^1, C_j^2, C_j^3)$ is a three-channel colour value. By projecting a 3D laser reading point $p_i = [x, y, z]$ with the range r_i, onto the image plane, a projected laser range reading point $\mathbf{R}_i = (X_i, Y_i, r_i, (C_i^1, C_i^2, C_i^3))$ is obtained, which associates a range value r_i with the coordinates and the colour of an image pixel.

The interpolation problem can now be stated for a given pixel $\mathcal{P}_j$ and a set of projected laser range readings $\mathbf{R} = \mathbf{R_i}$, as to estimate the interpolated range reading r_j^* as accurately as possible.

The visual feature f_i is located in the image at position $[X,Y]$. The depth estimate r_i^* is assigned to the laser range reading r_i corresponding to the projected laser range reading $\mathbf{R}_i$ which is closest (regarding pixel distance) to $[X,Y]$.

Note that there are more accurate methods which also incorporates visual information in the interpolation [1], however by utilizing the covariance of each visual feature point the depth error will have less impact in ambiguous cases.

C. Rigid Iterative Closest Point

The iterative closest points (ICP) algorithm [4, 5], finds the rigid body transformation between two scenes by minimizing the following constraint

$$J(\mathbb{R}, \mathbf{t}) = \sum_{i=1}^{N} ||y_i - \mathbb{R}x_i - \mathbf{t}||^2, \tag{1}$$

where x_i and y_i are the corresponding (closest) points from the different scenes. The selection of the corresponding pairs is done, in the standard version of ICP, by using a distance metric to search for the closest point. This search is the most time consuming part of the ICP algorithm. To decrease the search time a common approach is to create a kd-tree. Original ICP and other least squares methods assume identical and independent Gaussian noise.

To obtain the rigid transformation that minimizes the above equation, there exists various closed-form solutions. In our approach we have adopted the singular value decomposition method proposed by Arun et al. [2].

In our approach, the correspondences are detected using visual features, i.e. an exhaustive search for closest points is not required. In addition, since our method relies on a vision based approach the assumption of identical and independent noise of the feature point is a problematic approximation as discussed below.

D. Rigid Generalized Total Least Squares ICP

Generalized Total Least Square ICP (GTLS-ICP) has been proposed by San-Jose et. al. [15] as an extension of ICP. This method is similar to standard ICP but also incorporates a covariance matrix for each point. Instead of minimizing Eq. 1, GTLS-ICP utilizes the following function:

$$\begin{aligned} J(\mathbb{R}, \mathbf{t}) = &\sum_{i=1}^{N} (q_i - y_i)^T C_{q_i}^{-1} (q_i - y_i) + \\ &\sum_{i=1}^{N} (y_i - q_i)^T C_{y_i}^{-1} (y_i - q_i), \end{aligned} \tag{2}$$

where $q_i = \mathbb{R}x_i + \mathbf{t}$. The covariance matrix C_{q_i} is obtained by rotating the eigen vectors of the covariance matrix C_{x_i} with the rotation matrix $\mathbb{R}$. However, there is no closed-form solution to minimize this function and the method instead iteratively estimates the rigid body transformation $\mathbb{R}$ and $\mathbf{t}$. In our implementation we first use the standard ICP method (previous Section) and after convergence then apply a conjugate gradient method to minimize Eq. 2.

To obtain a covariance for each visual feature point, the closest projected laser point p_0 relative to the visual feature in the image plane, see Section III-B are used together with M surrounding laser points. The covariance C is calculated as

$$C = \frac{1}{M} \sum_{i=0}^{M} (p_i - \mu)^2, \tag{3}$$

where $\mu = \frac{1}{M+1} \sum_{i=0}^{M} p_i$. In our experimental evaluation $M = 8$, see Fig. 1.

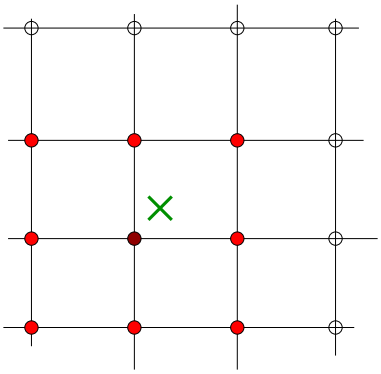

Fig. 1. Laser points used to estimate the covariance. The green cross ($\times$) represents a visual feature. Circles represent range readings, where filled red dot represent range readings used to obtain the covariance estimate. The central dot represents the laser point to determine the depth of the visual feature. The horisontal lines represent the 2D laser reading and the vertical lines the tilt movement of the wrist.

E. Rigid Trimmed Extension

Since visual features are used to establish corresponding scan points, no further means of data association, (such as searching for closest data points in ICP) is necessary. Although the SIFT features were found to be very discriminative (see for example [11]), there is of course still a risk that some of the correspondences are not correct. To further decrease the possibility of erroneous point associations, only a set fraction of the correspondences with the smallest spatial distance between corresponding points is used for registration. This, in addition, also removes points with the correct correspondances but with non-consistent depth estimate. In the experiments presented in this paper the fraction was set to 70%. Because the fraction of data points that is used to estimate the relative pose $[\mathbb{R}, \mathbf{t}]_t$ between two scans depends on the previous estimate $[\mathbb{R}, \mathbf{t}]_{t-1}$ (since the relative pose estimate affects the spatial distance between corresponding points), the minimization needs to be applied in an iterative manner. Thus relative pose updates are calculated repeatedly with the minimization using the previous estimate $[\mathbb{R}, \mathbf{t}]_{t-1}$ as input to the next iteration step until a stopping criterion is met. To obtain an inital pose estimate the 70% fraction of the pairs was randomly selected in the first iteration. However any initial pose estimate can be used. The suggested approach is similar to the RANSAC algorithm [6] applied directly in 3D (not using planar 2D image coordiantes), where the new model is determined directly with the closed form solution. One difference is that the suggested approach do not require a threshold value to determine inliers. As the stopping criterion in the experiments in this paper we used that if the change of the mean squared error (MSE) of the spatial distance between the corresponding points compared to the previous iteration was less than $10^{-6}\ m^2$.

Note that the spatial distance between corresponding points is used even if the covariance based ICP method is used to select the 70% fraction of the corresponding points. Otherwise points with high covariance, which have a small impact in Eq. 2 will more likely be selected in the trimmed version. By using a selection criterion based on the spatial distance we want to avoid a loss of performance, which was more notable in the cases where many (> 30) matches were available.

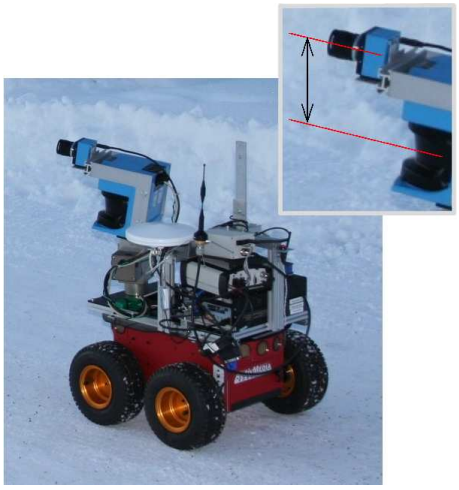

Fig. 2. Our mobile robot platform "Tjorven" equipped with the sensors used in this paper: the SICK LMS 200 laser range scanner and a colour CCD camera both mounted on an Amtec pan tilt unit. The close-up shows the displacement between the camera and the laser which causes parallax errors.

IV. Experimental Setup

A. Hardware

For the experiments presented in this paper we used the ActivMedia P3-AT robot "Tjorven" shown in Fig. 2, equipped with a 2D laser ranger scanner (SICK LMS 200) and a 1-MegaPixel (1280x960) colour CCD camera. The CCD camera and the laser scanner are both mounted on a pan-tilt unit from Amtec with a displacement between the optical axes of approx. 0.2 m. The angular resolution of the laser scanner was set to 0.25 degrees.

B. Data Collection

For each pose, 3D range and image data are collected as follows. First, three sweeps are carried out with the laser scanner at -60, 0 and 60 degrees relative to the robot orientation (horizontally). During each of these sweeps, the tilt of the laser scanner is continuously shifted from -40 degrees (looking up) to 30 degrees (looking down). After the three range scan sweeps, seven camera images are recorded at -90, -60, -30, 0, 30, 60, and 90 degrees relative to the robot orientation (horizontally) and at a fixed tilt angle of -5 degrees (looking up). The full data set acquired at a single scan pose is shown on Fig. 3.

C. Calibration

In our setup the displacement between the laser scanner and the camera is fixed. Thus it is necessary to determine 6 external calibration parameters (3 for rotation and 3 for translation) once. This is done by simultaneously optimizing the calibration parameters for several calibration scans. The method we use requires a special calibration board, see Fig. 4, which is also used to determine the internal calibration parameters of

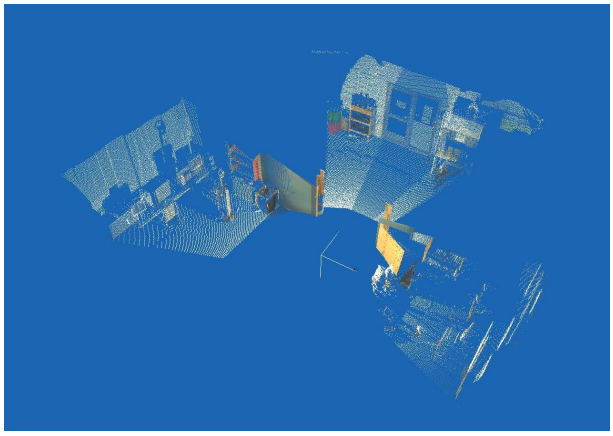

Fig. 3. Full data set acquired for a single scan pose comprising three sweeps with the laser scanner fused with colour information from seven camera images.

Fig. 4. Calibration board used to determine the calibration parameters of the camera, with a chess board texture and reflective tape (gray border) to locate the board using remission / intensity values from the laser scanner.

the camera. The calibration board is pasted with reflective tape at its borders enabling to use the reflective (remission) values from the laser scanner to automatically estimate the 3D position of the chess board corners detected in the image. The external parameters for the camera are obtained by minimizing the sum of squared distances (SSD) between the chess board corners found in the image and the 3D position of the chess board corners derived from the laser range readings.

D. Experiment

To evaluate the registration, a data set consisting of 22 scan poses, i.e. from 66 laser scanner sweeps and 154 camera images as described in Section IV-B was collected in an indoor lab environment. The first scan pose and the last scan pose were collected at a similar position. An example of registration result can be seen in Fig. 5.

The performance metric of the registration method is the translation and angular distance between the estimated pose from the registration method compared to the ground truth. Since the first and the last scan pose were taken at a similar position, the ground truth was determined by matching the first scan pose with the last scan pose using the trimmed ICP version using all available corresponding visual feature points. The estimated position was calculated by sequentially registeing all 22 scan poses, which means that only one small failure in one of the registrations will heavily influence the final pose estimate.

To better evaluate the registration method, the number of corresponding matches $\mathcal{N}$ that was used in the registration was also investigated together with the number of required iterations.

V. Results

Table I, II show the euclidean pose error d (in meters) together with the sum of the rotational error α (in radians). Since corresponding matches were done randomly, each sequential registration was repeated 5 times. These initial results show that GTLS-ICP works better when there are few corresponding matches and when the number of available matches increases the two methods show more similar results. The increased error with higher number of corresponding points $\mathcal{N}$ is likely to be caused by the random selection of points.

Table III shows the number of iterations required for convergence for the trimmed closed form ICP version. Note that in GTLS-ICP a conjugate gradient minimization method is applied.

TABLE I
Registration results $\mathcal{N} = [10, 15, 20]$, given in meters and radians using the trimmed registration versions

	$Tr.\ ICP$			$Tr.\ GTLS-ICP$		
$\mathcal{N}$	10	15	20	10	15	20
d	1.14	0.76	0.30	0.84	0.70	0.24
σ_d	0.54	0.83	0.11	0.33	0.85	0.14
α	0.30	0.17	0.05	0.25	0.18	0.06
σ_α	0.18	0.24	0.02	0.11	0.22	0.04

TABLE II
Registration results $\mathcal{N} = [30, 40, 60]$, given in meters and radians using the trimmed registration versions.

	$Tr.\ ICP$			$Tr.\ GTLS-ICP$		
$\mathcal{N}$	30	40	60	30	40	60
d	0.09	0.14	0.13	0.11	0.19	0.15
σ_d	0.05	0.07	0.03	0.08	0.10	0.06
α	0.04	0.03	0.03	0.04	0.04	0.03
σ_α	0.02	0.01	0.01	0.01	0.02	0.02

TABLE III
Number of iterations required for convergence using the trimmed version of the closed form ICP method

$\mathcal{N}$	10	15	20	30	40	60
$\#_{iter}$	4.46	4.95	5.20	5.00	6.00	5.97
σ_{iter}	0.68	0.86	0.93	1.87	1.07	1.07

VI. Conclusions and Future Work

In this paper we have suggested a vision based registration method that uses visual features to handle the correspondence problem. The method integrates both vision and a 3D laser

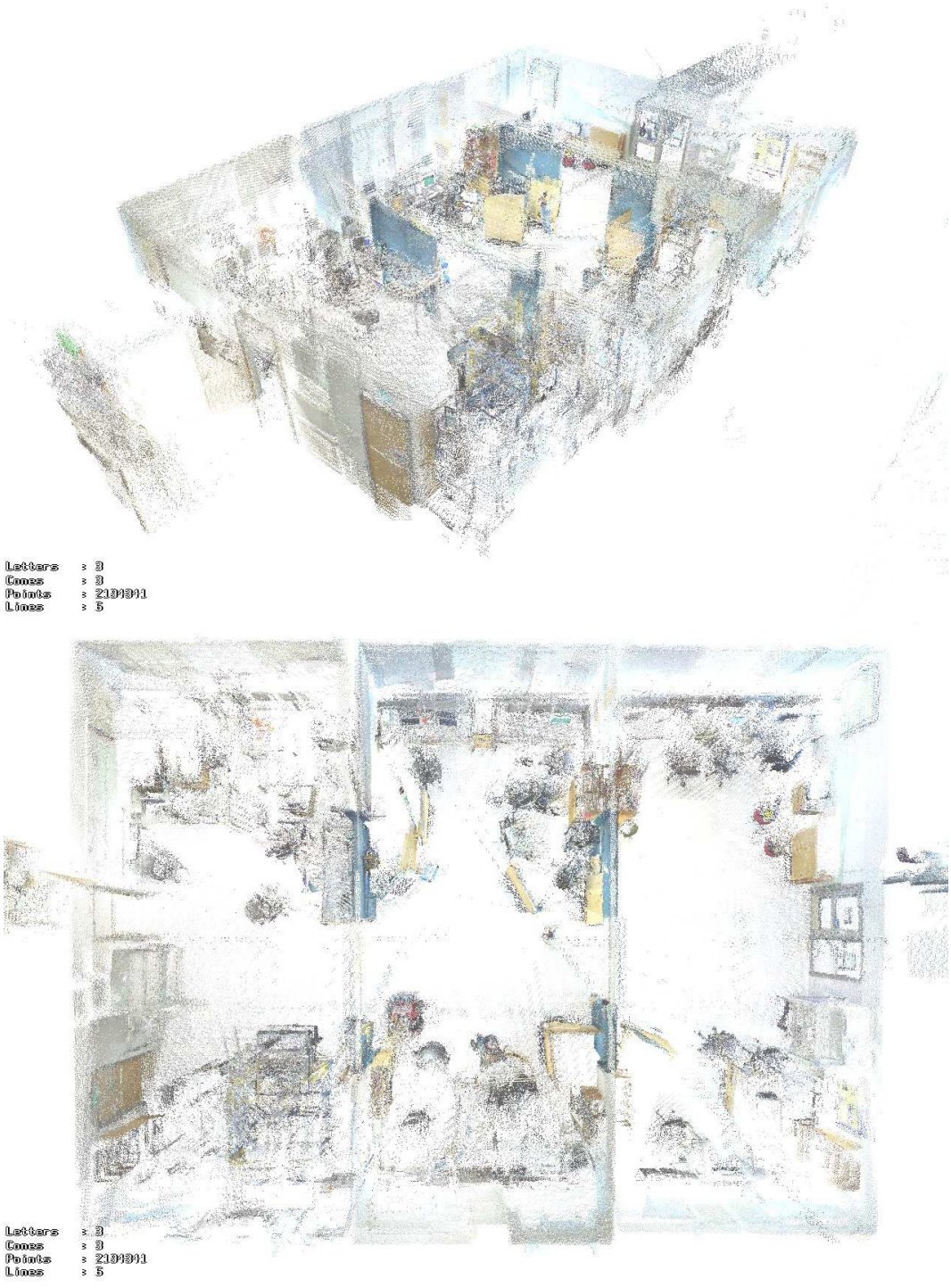

Fig. 5. A registration result generated by sequential registration of 22 scan poses. The visualized data consists of 3×22 registered scans and the corresponding colours from 7×22 camera images.

scanner and does not rely on any initial estimate. The 3D laser scanner is used to obtain a depth estimate and a covariance estimate of the extracted visual feature which is incorporated in the registration. An initial experiment has been conducted to verify the approach.

Our ongoing work includes a more thoroughly evaluation of the method and to test the method on more challenging data sets. Also to do a performance comparison with other "plain" laser based registration technique, such as ICP or iterative 3D-NDT [10].

REFERENCES

[1] Henrik Andreasson, Rudolph Triebel, and Achim Lilienthal. Vision-based interpolation of 3d laser scans. In *Proceedings of the 2006 IEEE International Conference on Autonomous Robots and Agents (ICARA 2006)*, pages 455–460, Palmerston North, New Zeeland, 2006. IEEE.

[2] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9(5):698–700, 1987.

[3] T.D. Barfoot. Online visual motion estimation using fastslam with sift features. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 579–585, 2005.

[4] P. J. Besl and N. D. McKay. A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[5] Y. Chen and G. Medioni. Object Modelling by Registration of Multiple Range Images. *Image and Vision Computing*, 10(3):145–155, 1992.

[6] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[7] J. Gonzalez-Barbosa and S. Lacroix. Rover localization in natural environments by indexing panoramic images. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pages 1365–1370, 2002.

[8] Kin Leong Ho and Paul Newman. Loop closure detection in slam by combining visual and spatial appearance. *Robotics and Autonomous System*, 54(9):740–749, September 2006.

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

[10] Martin Magnusson, Tom Duckett, and Achim Lilienthal. 3D Scan Registration for Autonomous Mining Vehicles. *Journal of Field Robotics*, page to appear, 2007.

[11] Krystian Mikolajczyk and Cordelia Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.

[12] Paul M. Newman, David M. Cole, and Kin Leong Ho. Outdoor SLAM using visual appearance and laser ranging. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pages 1180–1187, 2006.

[13] Andreas Nüchter, Kai Lingemann, Joachim Hertzberg, and Hartmut Surmann. Heuristic-based laser scan matching for outdoor 6d slam. In *KI*, pages 304–319, 2005.

[14] Daniel C. Asmar Samer M. Abdallah and John S. Zelek. Towards benchmarks for vision SLAM algorithms. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pages 1542–1547, 2006.

[15] R. San-Jose, A. Brun, and C.-F. Westin. Robust generalized total least squares iterative closest point registration. In *Seventh International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'04)*, Lecture Notes in Computer Science, Rennes - Saint Malo, France, September 2004.

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

[17] Rudolph Triebel, Patrick Pfaff, and Wolfram Burgard. Multi-level surface maps for outdoor terrain mapping and loop closing. In *"Proc. of the International Conference on Intelligent Robots and Systems (IROS)"*, pages 2276–2282, 2006.