Automatic Appearance-Based Loop Detection from Three-Dimensional Laser Data Using the Normal Distributions Transform

Martin Magnusson and Henrik Andreasson
Center for Applied Autonomous Sensor Systems
Örebro University
70182 Örebro, Sweden
e-mail: martin.magnusson@oru.se,
henrik.andreasson@oru.se

Andreas Nüchter
School of Engineering and Science
Jacobs University Bremen
28759 Bremen, Germany
e-mail: andreas@nuechti.de

Achim J. Lilienthal
Center for Applied Autonomous Sensor Systems
Örebro University
70182 Örebro, Sweden
e-mail: achim@lilienthals.de

Received 5 January 2009; accepted 5 August 2009

We propose a new approach to appearance-based loop detection for mobile robots, using three-dimensional (3D) laser scans. Loop detection is an important problem in the simultaneous localization and mapping (SLAM) domain, and, because it can be seen as the problem of recognizing previously visited places, it is an example of the data association problem. Without a flat-floor assumption, two-dimensional laser-based approaches are bound to fail in many cases. Two of the problems with 3D approaches that we address in this paper are how to handle the greatly increased amount of data and how to efficiently obtain invariance to 3D rotations. We present a compact representation of 3D point clouds that is still discriminative enough to detect loop closures without false positives (i.e., detecting loop closure where there is none). A low false-positive rate is very important because wrong data association could have disastrous consequences in a SLAM algorithm. Our approach uses only the appearance of 3D point clouds to detect loops and requires no pose information. We exploit the normal distributions transform surface representation to create feature histograms based on surface orientation and smoothness. The surface shape histograms compress the input data by two to three orders of magnitude. Because of the high compression rate, the histograms can be matched efficiently to compare the appearance of two scans. Rotation invariance is achieved by aligning scans with respect to dominant surface orientations. We also propose to use expectation maximization to fit

Journal of Field Robotics 26(11–12), 892–914 (2009)
Published online in Wiley InterScience (www.interscience.wiley.com). • DOI: 10.1002/rob.20314

a gamma mixture model to the output similarity measures in order to automatically determine the threshold that separates scans at loop closures from nonoverlapping ones. We discuss the problem of determining ground truth in the context of loop detection and the difficulties in comparing the results of the few available methods based on range information. Furthermore, we present quantitative performance evaluations using three real-world data sets, one of which is highly self-similar, showing that the proposed method achieves high recall rates (percentage of correctly identified loop closures) at low false-positive rates in environments with different characteristics.

1. INTRODUCTION

For autonomously navigating mobile robots, it is essential to be able to detect when a loop has been closed by recognizing a previously visited place. One example application is when performing simultaneous localization and mapping (SLAM). A common way to perform SLAM is to let a robot move around in the environment, sensing its surroundings as it goes. Typically, discrete two-dimensional (2D) or three-dimensional (3D) laser scans are registered using a local scan registration algorithm in order to correct the robot's odometry and improve the estimate of the robot's pose (that is, its position and orientation) at each point in time. The scans can be stitched together at their estimated poses in order to build a map. However, even though good scan registration algorithms exist [for example, 3D-normal distributions transform (NDT) (Magnusson, Lilienthal, & Duckett, 2007)], pose errors will inevitably accumulate over longer distances, and after covering a long trajectory the robot's pose estimate may be far from the true pose. When a loop has been closed and the robot is aware that it has returned to a previously visited place, existing algorithms can be used to distribute the accumulated pose error of the pairwise registered scans in order to render a consistent map. Some examples are the tree-based relaxation methods of Frese, Larsson, and Duckett (2005) and Frese and Schröder (2006) and the 3D relaxation methods of Grisetti, Grzonka, Stachniss, Pfaff, and Burgard (2007) and Borrmann, Elseberg, Lingemann, Nüchter, and Hertzberg (2008). However, *detecting* loop closure when faced with large pose errors remains a difficult problem. As the uncertainty of the estimated pose of the robot grows, an independent means of detecting loop closure becomes increasingly important. Given two 3D scans, the question we are asking is, "Have I seen this before?" A good loop detection algorithm aims at maximizing the recall rate; that is, the percentage of true positives (scans acquired at the same place that are recognized as such), with a minimum of false positives (scans that are erroneously considered to be acquired at the same place). False positives are much more costly than false negatives in the context of SLAM. A single false positive can render the map unusable if no further measures are taken to recover from false scan correspondences. On the other hand, a relatively low number of true positives is often acceptable, given that several scans are acquired from each overlapping section. As long as a few of these scans are detected, the loop can be closed.

We present a loop detection approach based on the appearance of scans. Appearance-based approaches often use camera images (Booij, Terwijn, Zivkovic, & Kröse, 2007; Cummins & Newman, 2007, 2008a, 2008b, 2009; Konolige et al., 2009; Valgren & Lilienthal, 2007). In this work, however, we consider only data from a 3D laser range scanner. Using the proposed approach, loop detection is achieved by comparing histograms computed from surface shape. The surface shape histograms can be used to recognize scans from the same location without pose information, thereby helping to solve the problem of global localization. Scans at loop closure are separated from nonoverlapping scans based on a "difference threshold" in appearance space. Pose estimates from odometry or scan registration are not required. (However, if such information is available, it could be used to further increase the performance of the loop detection by restricting the search space.) Though we have chosen to term the problem "loop detection," the proposed method solves the same problem that Cummins and Newman (2009) refer to as "appearance-only SLAM."

Existing 2D loop detection algorithms could potentially be used for the same purpose, after extracting a single scan plane from the available 3D scans. However, in many areas it may be advantageous to use all of the available information. One example is for vehicles driving over rough surfaces. Depending on the local slope of the surface, 2D scans from nearby

positions may look quite different. Therefore the appearance of 2D scans cannot be used to detect loop closure in such cases. In fact, this is a common problem for current 2D-scanning semi-autonomous mining vehicles.[1] Places where there are nearly horizontal surfaces close to the level of the 2D laser scanner are especially problematic. Another example is places where there are deep wheel tracks, meaning that a small lateral offset can result in a large difference in the vehicle's roll angle. Using 3D data instead of 2D is of course also important for airborne robots whose orientation is not restricted to a mainly planar alignment.

The work described in this paper builds on previously published results (Magnusson, Andreasson, Nüchter, & Lilienthal, 2009), with the main additions being a sound method for estimating the difference threshold parameter, a more complete performance evaluation, and more experimental data, as well as a discussion of the difficulties of determining ground truth for loop detection.

The paper is laid out as follows. In Section 2, we describe the details of our loop detection approach. The performance is evaluated in Section 3, and a method for automatically selecting the difference threshold is described and evaluated in Section 3.4. Our approach is compared to related work on loop detection in Section 4. Finally, the paper is summarized in Section 5, which also states our conclusions, and directions for future work are suggested in Section 6.

2. SURFACE SHAPE HISTOGRAMS

Our loop detection method is inspired by NDT. NDT is a method for representing a scan surface as a piecewise continuous and twice-differentiable function. It has previously been used for efficient pairwise 2D and 3D scan registration (Biber & Straßer, 2003; Magnusson et al., 2007; Magnusson, Nüchter, Lörken, Lilienthal, & Hertzberg, 2009). However, the NDT surface representation can also be used to describe the appearance of a 3D scan, as will be explained in this section.

2.1. The NDT

The NDT gives a compact surface shape representation, and it therefore lends itself to describing the general appearance of a location. The method is outlined in the following. For more details, refer to previous publications (Biber & Straßer, 2003; Magnusson et al., 2007).

Given a range scan represented as a point cloud, the space occupied by the scan is subdivided into a regular grid of cells (squares in the 2D case, cubes in the 3D case). Each cell c_i stores the mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ of the positions of the scan points within the cell; in other words, the parameters of a normally distributed probability density function (PDF) describing the local surface shape. Depending on the covariance, the PDF can take on a linear (stretched ellipsoid), planar (squashed ellipsoid), or spherical shape. Our appearance descriptor is created from histograms of these local shape descriptions.

To minimize the issues with spatial discretization we use overlapping cells, so that if the side length of each cell is q, the distance between each cell's center point is $q/2$. (The parameter choices will be covered in Section 2.5.)

2.2. Appearance Descriptor

It is possible to use the shapes of the PDF of NDT cells to describe the appearance of a 3D scan, classifying the PDFs based on their orientation and shape. For each cell, the eigenvalues $\lambda_1 \leq \lambda_2 \leq \lambda_3$ and corresponding eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ of the covariance matrix are computed. We use three main cell classes: spherical, planar, and linear. Distributions are assigned to a class based on the relations between their eigenvalues with respect to a threshold $t_e \in [0, 1]$ that quantizes a "much smaller" relation:

- Distributions are linear if $\lambda_2/\lambda_3 \leq t_e$.
- Distributions are planar if they are nonlinear and $\lambda_1/\lambda_2 \leq t_e$.
- Distributions are spherical if they are nonlinear and nonplanar (in other words, if no eigenvalue is $1/t_e$ times larger than another one).

The planar cell classes can be divided into subclasses, based on surface orientation. As stated in our earlier work (Magnusson, Andreasson, et al., 2009), the same can potentially be done for the linear classes, and spherical subclasses could be defined by surface roughness. However, in our experiments, using one spherical and one linear class has been

[1]This information is from personal communication with Johan Larsson, Atlas Copco Rock Drills.

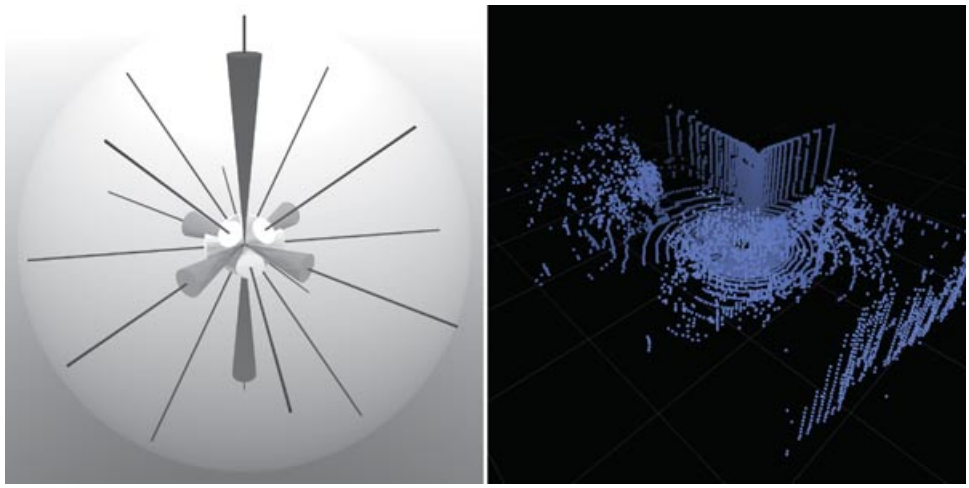


Figure 1. Visualization of the planar part $\boldsymbol{p}$ of a histogram vector created from the scan on the right. In this case, $p = 9$ planar directions are used. The thin black lines correspond to the directions $P_1, \ldots, P_9$. The cones are scaled according to the values of the corresponding histogram bins. There are two cones for each direction in this illustration: one on each side of the origin. The dominant directions used to normalize the scan's orientation are shaded. (The following text will be explained further in Section 2.3.) Directions that are not in $\mathcal{Z}$ or $\mathcal{Y}$ are white. $\mathcal{Z}$ (dark gray) contains one direction in this case: the vertical direction, corresponding to the ground plane. $\mathcal{Y}$ (light gray) includes two potential secondary peaks (whose magnitudes are more similar to the rest of the binned directions). In this example t_a was set to 0.6.

sufficient. It would also be straightforward to use more classes such as different levels of "almost planar" distributions by using more than one eigenvalue ratio threshold. However, for the data presented here, using more than one threshold t_e did not improve the result.

For the planar distributions, the eigenvector $\boldsymbol{e}_1$ (which corresponds to the smallest eigenvalue) coincides with the normal vector of the plane that is approximated by the PDF. We define planar subclasses as follows. Assuming that there is a set $\mathcal{P}$ of p approximately evenly distributed lines passing through the origin, $\mathcal{P} = \{P_1, \ldots, P_p\}$, the index for planar subclasses is

$$i = \arg\min_j \delta(\boldsymbol{e}_1, P_j), \tag{1}$$

where $\delta(\boldsymbol{e}, P)$ is the distance between a point $\boldsymbol{e}$ and a line P. In other words, we choose the index of the line P_j that is closest to $\boldsymbol{e}_1$. The problem of evenly distributing a number of lines intersecting the origin is analogous to distributing points evenly on the surface of a sphere. This is an ill-posed problem because it is not possible to find a solution in which the distances between all neighboring points are equal. However, a number of solutions giving approximately even point distributions exist. For example, using an equal area partitioning (Saff & Kuijlaars, 1997) to distribute p points on a half-sphere, $\mathcal{P}$ is the set of lines connecting the origin and one of the points. The distribution of lines used in this work is visualized in Figure 1.

Using p planar subclasses, the basic element of the proposed appearance descriptor is the feature vector

$$\boldsymbol{f} = \left[\underbrace{f_1, \ldots, f_p}_{\text{planar classes}}, \underbrace{f_{p+1}}_{\text{spherical}}, \underbrace{f_{p+2}}_{\text{linear}} \right]^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{p} \\ f_{p+1} \\ f_{p+2} \end{bmatrix}, \tag{2}$$

where f_i is the number of NDT cells that belong to class i.

In addition to surface shape and orientation, the distance from the scanner location to a particular surface is also important information. For this reason, each scan is described by a matrix

$$\mathbf{F} = \begin{bmatrix} \boldsymbol{f}_1 \cdots \boldsymbol{f}_r \end{bmatrix} \tag{3}$$

and a corresponding set of range intervals $\mathcal{R} = \{r_1, \ldots, r_r\}$. The matrix is a collection of surface shape histograms, where each column $\boldsymbol{f}_k$ is the histogram of all NDT cells within range interval r_k (measured from the laser scanner position).

2.3. Rotation Invariance

Because the appearance descriptor (3) explicitly uses the orientation of surfaces, it is not rotation invariant. For the appearance descriptor to be invariant to rotation, the orientation of the scan must first be normalized.

Starting from an initial histogram vector $\boldsymbol{f}'$, with a single range interval $\mathcal{R} = \{[0, \infty)\}$, we want to find two peaks in plane orientations and orient the scan so that the most common plane normal (the primary peak) is aligned along the z axis and the second most common (the secondary peak) is in the yz plane. The reason for using plane orientations instead of line orientations is that planar cells are much more common than linear ones. For an environment with more linear structures than planar ones, line orientations could be used instead, although such environments are unlikely to be encountered.

Because there is not always an unambiguous maximum, we use two *sets* of directions, $\mathcal{Z}$ and $\mathcal{Y}$. Given the planar part $\boldsymbol{p}' = [p'_1, \ldots, p'_p]^{\mathrm{T}}$ of $\boldsymbol{f}'$ and an ambiguity threshold $t_a \in [0, 1]$ that determines which histogram peaks are "similar enough," the dominant directions are selected as follows. (This selection is also illustrated in Figure 1.) First we pick the histogram bin with the maximum value

$$i' = \underset{i}{\operatorname{argmax}}\, p'_i. \tag{4}$$

The potential primary peaks are i' and any directions that are "almost" as common as i' with respect to t_a:

$$\mathcal{Z} = \{i \in \{1, \ldots, p\} \mid p'_i \geq t_a p'_{i'}\}. \tag{5}$$

The same procedure is repeated to find the second most common direction, choosing as a secondary peak the largest histogram bin that is not already included in the primary peak set:

$$i'' = \underset{i}{\operatorname{argmax}}\, p'_i \mid i \notin \mathcal{Z}. \tag{6}$$

The potential secondary peaks are i'' and any directions that are almost as common as i'' (but not already included in the primary peak set):

$$\mathcal{Y} = \{i \in \{1, \ldots, p\} \mid i \notin \mathcal{Z},\ p'_i \geq t_a p'_{i''}\}. \tag{7}$$

This procedure gives two disjoint subsets $\mathcal{Z} \subset \mathcal{P}$ and $\mathcal{Y} \subset \mathcal{P}$.

Now we want to align the primary (most common) peak along the positive z axis. In the following, we will use axis/angle notation for rotations: $\mathbf{R} = (\boldsymbol{v}, \phi)$ denotes a rotation with angle ϕ around the axis $\boldsymbol{v}$. To perform the desired alignment, the rotation

$$\mathbf{R}_z = \left(\boldsymbol{P}_i \times \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},\ -\arccos\left(\boldsymbol{P}_i \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) \right), \tag{8}$$

where $\boldsymbol{P}_i$ is a unit vector along the line P_i, rotates the scan so that $\boldsymbol{P}_i$ is aligned along the positive z axis. The rotation axis $\boldsymbol{P}_i \times [0, 0, 1]$ is perpendicular to the z axis. A separate rotation is created for each potential primary peak $i \in \mathcal{Z}$.

Similarly, for each secondary peak $i \in \mathcal{Y}$, it is possible to create a rotation $\mathbf{R}_y$ that rotates the scan around the z axis so that $\boldsymbol{P}_i$ lies in the yz plane. To determine the angle of $\mathbf{R}_y$, we use the normalized projection of $\mathbf{R}_z \boldsymbol{P}_i$ onto the xy plane: $\boldsymbol{P}'_i$. The angle of $\mathbf{R}_y$ is the angle between the projected vector $\boldsymbol{P}'_i$ and the yz plane:

$$\mathbf{R}_y = \left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},\ -\arccos\left(\boldsymbol{P}'_i \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right) \right). \tag{9}$$

Given a scan $\mathcal{S}$, the appearance descriptor $\mathbf{F}$ is created from the rotated scan $\mathbf{R}_y \mathbf{R}_z \mathcal{S}$. This alignment is always possible to do, unless all planes have the same orientation. If it is not possible to find two main directions, it is sufficient to use only $\mathbf{R}_z$ because in that case no subsequent rotation around the z axis changes which histogram bins are updated for any planar PDF. If linear subclasses of different orientations are used, it is possible to derive $\mathbf{R}_y$ from linear directions if only one planar direction can be found.

In the case of ambiguous peaks (that is, when $\mathcal{Z}$ or $\mathcal{Y}$ has more than one member), we generate multiple histograms. For each combination $\{i, j \mid i \in \mathcal{Z},\ j \in \mathcal{Z} \cup \mathcal{Y},\ i \neq j\}$ we apply the rotation $\mathbf{R}_y \mathbf{R}_z$ to the original scan and generate a histogram. The outcome is a set of histograms

$$\mathcal{F} = \{\mathbf{F}_1, \ldots, \mathbf{F}_{|\mathcal{Z}|(|\mathcal{Z} \cup \mathcal{Y}|-1)}\}. \tag{10}$$

The set $\mathcal{F}$ is the appearance descriptor of the scan.

For highly symmetrical scans, the approach presented in this section could lead to a very large number of histograms. For example, in the case of a scan generated at the center of a sphere, where the histogram bins for all directions have the same value, $p^2 - p$ histograms would be created (although a post-processing step to prune all equivalent histograms could reduce this to just one). In practice, this kind of symmetry effect has not been found to be a problem. The average number of histograms per scan is around three for the data sets used in this work.

2.4. Difference Measure

To quantify the difference between two surface shape histograms $\mathbf{F}$ and $\mathbf{G}$, we normalize $\mathbf{F}$ and $\mathbf{G}$ with their entrywise 1-norms (which corresponds to the number of occupied NDT cells in each scan), compute the sum of Euclidean distances between each of their columns (each column corresponds to one range interval), and weight the sum by the ratio $\max(\|\mathbf{F}\|_1, \|\mathbf{G}\|_1)/\min(\|\mathbf{F}\|_1, \|\mathbf{G}\|_1)$:

$$\sigma(\mathbf{F}, \mathbf{G}) = \sum_{i=1}^{r} \left(\left\| \frac{\boldsymbol{f}_i}{\|\mathbf{F}\|_1} - \frac{\boldsymbol{g}_i}{\|\mathbf{G}\|_1} \right\|_2 \right) \frac{\max(\|\mathbf{F}\|_1, \|\mathbf{G}\|_1)}{\min(\|\mathbf{F}\|_1, \|\mathbf{G}\|_1)}. \tag{11}$$

The normalization makes it possible to use a single threshold for data sets that contain both scans that cover a large area (with many occupied NDT cells) and scans of more confined spaces (with fewer cells). If the Euclidean distance without normalization were used instead,

$$\sigma(\mathbf{F}, \mathbf{G}) = \sum_{i=1}^{r} \left\| \boldsymbol{f}_i - \boldsymbol{g}_i \right\|_2, \tag{12}$$

scans with many cells would tend to have larger difference values than scans with few cells. A consequence is that in environments with some narrow passages and some open areas, the open spaces would be harder to recognize. It would not be possible to use a global, fixed threshold, because the best threshold for the wide areas would tend to cause false positives in the narrow areas.

The scaling factor $\max(\|\mathbf{F}\|_1, \|\mathbf{G}\|_1)/\min(\|\mathbf{F}\|_1, \|\mathbf{G}\|_1)$ is used to differentiate large scans (with many cells) from small ones (with few cells).

Given two scans $\mathcal{S}_1$ and $\mathcal{S}_2$ with histogram sets $\mathcal{F}$ and $\mathcal{G}$, all members of the scans' sets of histograms are compared to each other using Eq. (11), and the minimum σ is used as the difference measure for the scan pair:

$$\tau(\mathcal{S}_1, \mathcal{S}_2) = \min_{i,j} \sigma(\mathbf{F}_i, \mathbf{G}_j), \qquad \mathbf{F}_i \in \mathcal{F}, \mathbf{G}_j \in \mathcal{G}. \tag{13}$$

If $\tau(\mathcal{S}_1, \mathcal{S}_2)$ is less than a certain difference threshold value t_d, the scans $\mathcal{S}_1$ and $\mathcal{S}_2$ are assumed to be from the same location. For evaluation purposes the two scans $\mathcal{S}_1$ and $\mathcal{S}_2$ are classified as *positive*.

2.5. Parameters

Summarizing the preceding text, these are the parameters of the proposed appearance descriptor along with the parameter values selected for the experiments:

- NDT cell size $q = 0.5$ m
- range limits $\mathcal{R} = \{[0, 3), [3, 6), [6, 9), [9, 15), [15, \infty)\}$ m
- planar class count $p = 9$
- eigenvalue ratio threshold $t_e = 0.10$
- ambiguity ratio threshold $t_a = 0.60$

The values of these parameters were chosen empirically. Some parameters depend on the scale of the environment, but a single parameter set worked well for all investigated data sets.

The best cell size q and the range limits $\mathcal{R}$ depend mainly on the scanner configuration. If the cell size is too small, the PDFs are dominated by scanner noise. For example, planes at the farther parts of scans (where scan points are sparse) may show up in the histogram as lines with varying orientations. If the cell size is too large, details are lost because the PDFs do not accurately represent the surfaces. Previous work (Magnusson et al., 2007) has shown that cell sizes between 0.5 and 2 m work well for registering scans of the scale encountered by a mobile robot equipped with a rotating SICK LMS 200 laser scanner when using NDT for scan registration. We have used similar experimental platforms for the data examined in this work. For the present experiments, $q = 0.5$ m and $\mathcal{R} = \{[0, 3), [3, 6), [6, 9), [9, 15), [15, \infty)\}$ were used. Using fewer range intervals decreased the loop detection accuracy. If using a scanner with a different max range, $\mathcal{R}$ and q should probably be adjusted. The same parameter settings worked well for all the data sets

used here even though the point cloud resolution varies, with almost an order of magnitude among them.

Using nine planar classes (in addition to one spherical class and one linear class) worked well for all of the data sets. The reason for using only one spherical and linear class is that these classes tend to be less stable than planar ones. Linear distributions with unpredictable directions tend to occur at the far ends of a scan, where the point density is too small. Spherical distributions often occur at corners and edges, depending on where the boundaries of the NDT cells end up, and may shift from scan to scan. However, using the planar features only decreased the obtainable recall rate without false positives with around one-third for the data sets used in our evaluations. The small number of classes means that the surface shape histograms provide a very compact representation of the input data. We use only 55 values (11 shape classes and 5 range intervals) for each histogram. To achieve rotation invariance we use multiple histograms per scan, as described in Section 2.3, but with an average of three histograms per scan, the appearance of a point cloud with several tens of thousands of points can still be represented using only 165 values.

The eigenvalue ratio threshold t_e and ambiguity ratio threshold t_a were also chosen empirically. Both of these thresholds must be on the interval $[0, 1]$. In the experiments, using $t_e = 0.10$ and $t_a = 0.60$ produced good results independent of the data.

In addition to the parameters of the appearance descriptor, it is necessary to select a difference threshold t_d that determines which scans are similar enough to be assumed to have been taken at the same location. The difference threshold t_d was chosen separately for each data set, as described in Section 3.3. A method for automatically selecting a difference threshold is presented in Section 3.4.

3. EXPERIMENTS

3.1. Data Sets

To evaluate the performance of the proposed loop detection algorithm, three data sets were used: one outdoor set from a campus area, one from an indoor office environment, and one from an underground mine. All of the data sets are available online from the Osnabrück Robotic 3D Scan Repository (http://kos.informatik.uni-osnabrueck.de/3Dscans/).

The Hannover2 data set, shown in Figure 2, was recorded by Oliver Wulf at the campus of Leibniz Universität Hannover. It contains 922 3D omniscans (with 360-deg field of view) and covers a trajectory of about 1.24 km. Each 3D scan is a point cloud containing approximately 15,000 scan points. Ground truth pose measurements were acquired by registering every 3D scan against a point cloud made from a given 2D map and an aerial LIDAR scan made while flying over the campus area, as described in the SLAM benchmarking paper by Wulf, Nüchter, Hertzberg, and Wagner (2007).

The AASS-loop data set was recorded around the robot lab and coffee room of the Center for Applied Autonomous Sensor Systems (AASS) research institute at Örebro University. An overhead view of this data set is shown in Figure 3. The total trajectory traveled is 111 m. This set is much smaller than the Hannover2 one. It contains 60 omnidirectional scans with around 112,000 points per scan. For this data set, pairwise scan registration using 3D-NDT (given the initial pose estimates from the robot's odometry) was exact enough to be used for the ground truth poses. (The accumulated pose error between scan 1 and scan 60 was 0.67 m and 1.3 deg after registration.) However, using only the laser scans without odometry information, it is not possible to detect loop closure with NDT scan registration.

A third data set, Kvarntorp, was recorded in the Kvarntorp mine outside Örebro, Sweden. The original data set is divided into four "missions." For the experiments presented here, we used "mission 4" followed by "mission 1." The reason for choosing these two missions is that they overlap each other and that the starting point of mission 1 is close to the end point of mission 4. This combined sequence has 131 scans, each covering a 180-deg horizontal field of view and containing around 70,000 data points. The total trajectory is approximately 370 m. See Figure 4 for an overview of this data set. The Kvarntorp data set is rather challenging for a number of reasons. First, the mine environment is highly self-similar. Without knowledge of the robot's trajectory, it is very difficult to tell different tunnels apart, both from 3D scans and from camera images, as illustrated in Figure 5. The fact that the scans of this data set are not omnidirectional also makes loop detection more difficult, because the same location can look very different depending on which direction the scanner is pointing toward. Another challenge is that the distance traveled between the scans is longer for this data set. For

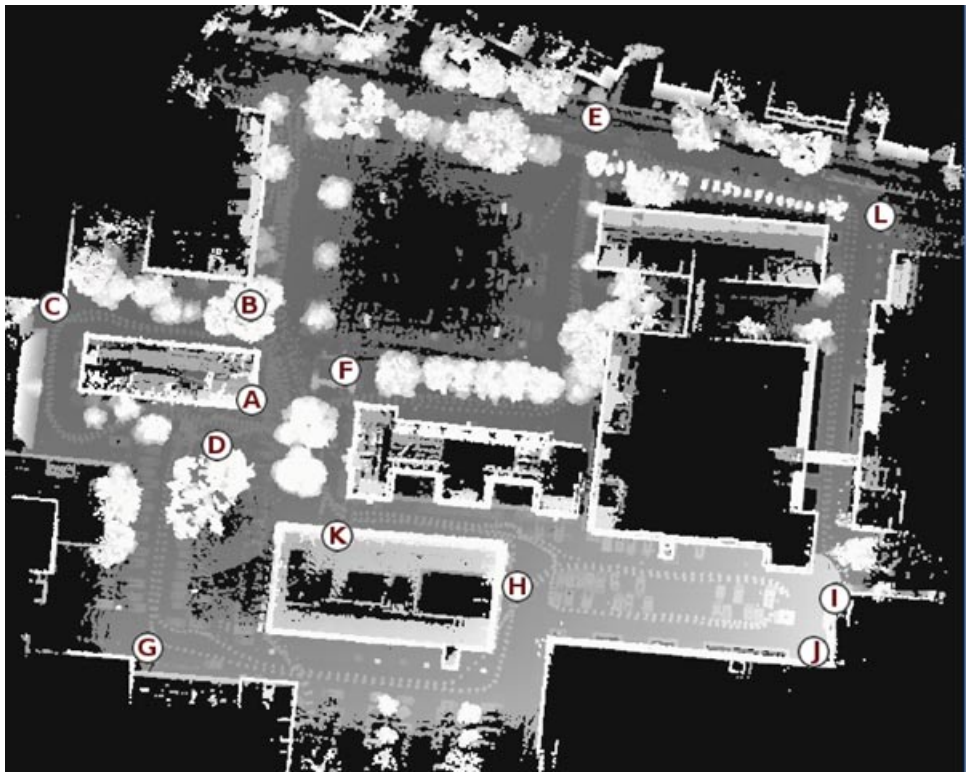


Figure 2. The Hannover2 data set, seen from above with parallel projection.

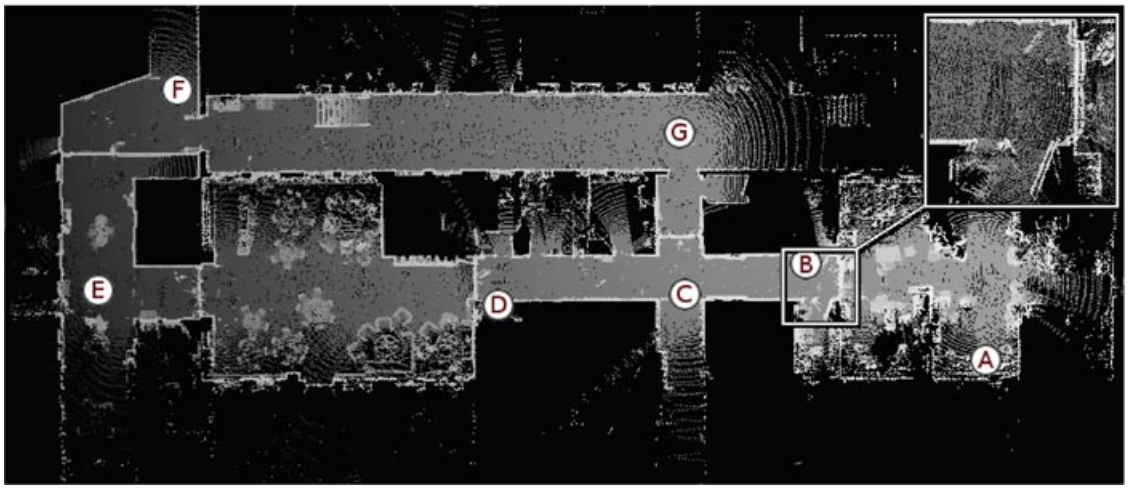


Figure 3. The AASS-loop data set, shown from above with the ceiling removed. The inlay in the right-hand corner shows the accumulated pose error from pairwise scan registration using 3D-NDT when returning to location B.

Figure 4. The Kvarntorp data set, seen from above with the ceiling removed.

this reason, scans taken when revisiting a location tend to be recorded farther apart, making the scans look more different.

Scan registration alone was not enough to build a consistent 3D map of the Kvarntorp data set, and an aerial reference scan was not available for obvious reasons. Instead, ground truth poses were provided using a network-based global relaxation method for 3D laser scans (Borrmann et al., 2008). A network with loop closures was manually created and given as input to the algorithm in order to generate a reference map. The result was visually inspected for correctness. The relaxation method of Borrman et al. will be briefly described here. Given a network with $n+1$ nodes $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n$ representing the poses $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_n$ and the directed edges $\boldsymbol{d}_{i,j}$, the algorithm aims at estimating all poses optimally. The directed edge $\boldsymbol{d}_{i,j}$ represents the *change* of the pose $(x, y, z, \theta_x, \theta_y, \theta_z)$ that is necessary to transform one pose $\boldsymbol{v}_i$ into $\boldsymbol{v}_j$; that is, $\boldsymbol{v}_i = \boldsymbol{v}_j \oplus \boldsymbol{d}_{i,j}$, thus transforming two nodes of the graph. For simplicity, the approximation that the measurement equation is linear is made; that is,

$$\boldsymbol{d}_{i,j} = \boldsymbol{x}_i - \boldsymbol{x}_j. \tag{14}$$

A detailed derivation of the linearization is given in the paper by Borrman et al. (2008). An error function is formed such that minimization results in improved pose estimations:

$$\mathbf{W} = \sum_{(i,j)} (\boldsymbol{d}_{i,j} - \bar{\boldsymbol{d}}_{i,j})^{\mathrm{T}} \mathbf{C}_{i,j}^{-1} (\boldsymbol{d}_{i,j} - \bar{\boldsymbol{d}}_{i,j}), \tag{15}$$

where $\bar{\boldsymbol{d}}_{i,j} = \boldsymbol{d}_{i,j} + \Delta\boldsymbol{d}_{i,j}$ models random Gaussian noise added to the unknown exact pose $\boldsymbol{d}_{i,j}$. This representation involves resolving the nonlinearities resulting from the additional roll and pitch angles by Taylor expansion. The covariance matrices $\mathbf{C}_{i,j}$ describing the pose relations in the network are computed based on the paired closest points. The error function (15) has a quadratic form and is therefore solved in closed form by sparse Cholesky decomposition.

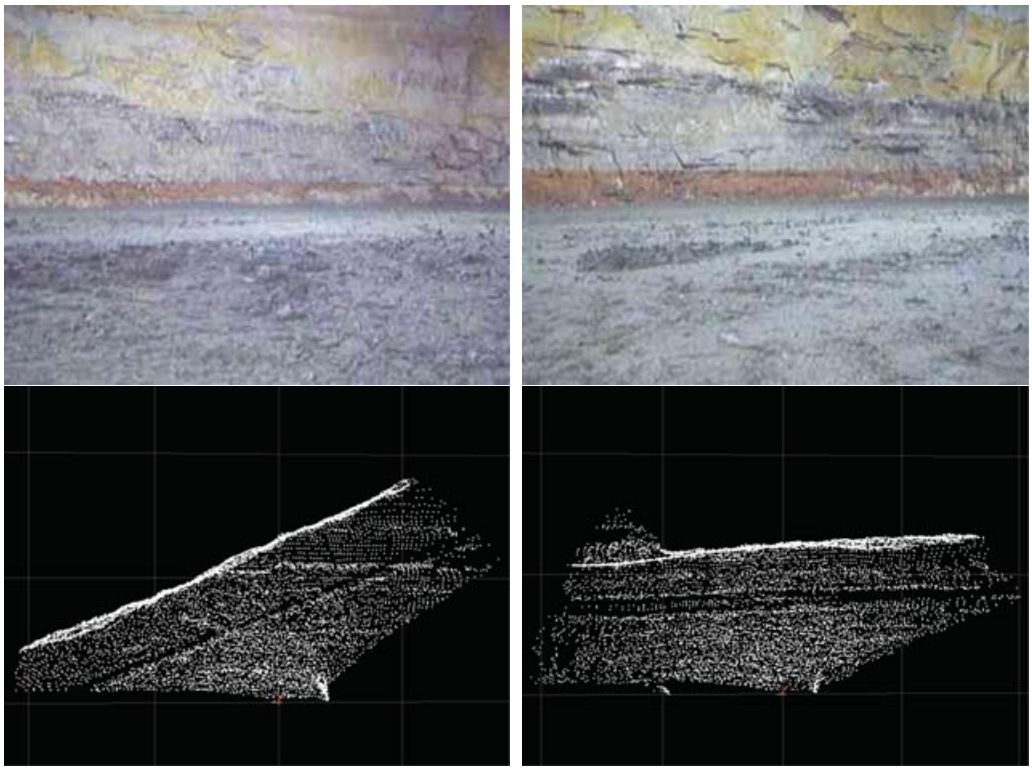
3.2. Experimental Method

We have used two methods to judge the discrimination ability of our surface shape histograms.

3.2.1. Full Evaluation

First, we look at all combinations $(\mathcal{S}_i, \mathcal{S}_j \mid i \neq j)$ of scan pairs from each data set, counting the number of true positives and false positives with regard to the ground truth. In related work (Bosse & Zlot, 2008b; Granström, Callmer, Ramos, & Nieto, 2009), the performance is reported as the recall rate with a manually chosen threshold that gives a 1% false-positive rate. For these tests, we have taken the same approach to evaluate the result.

However, it is not trivial to determine the ground truth: what should be considered a true or a false positive. In this work we have chosen to use the matrix of the distances between all scan pairs as the ground truth, after applying a distance threshold t_r (in metric space), so that all pairs of scans that are within, for example, 3 m are considered to be truly overlapping. It is not always easy to select a distance threshold value that captures the relationships between scans in a satisfactory manner. If the threshold t_r is large, some scans with very different appearances (for example, scans taken at different sides of the corner of a building or before and after passing through a door) might still be considered to overlap and will therefore be regarded as false negatives when their appearances do not match. Another problem is that sequential scans are often acquired in close proximity



(a) Location F (b) Location H

Figure 5. An example of perceptual aliasing in the Kvarntorp data set. The images show two different places (locations F and H in Figure 4). It is difficult to tell the two places apart, from both the camera images and the scanned point clouds. The point clouds are viewed from above.

to one another. Therefore, when revisiting a location, there will be several overlapping scans within the distance threshold, according to the "ground truth." But with a discriminative difference threshold t_d (in appearance space), only one or a few of them may be detected as positives. Even when the closest scan pair is correctly matched, the rest would then be regarded as false negatives, which may not be the desired result. If, on the other hand, the distance threshold t_r is too small, the ground truth matrix will miss some loop closures where the robot is not revisiting the exact same position.

Another possibility would be to manually label all scan pairs. However, when evaluating multiple data sets containing several hundreds of scans, it is not practical to do so; and even then, some arbitrary decision would have to be made as to whether some scan pairs overlap.

We will discuss how our experimental method compares to the evaluations of other authors in Section 4. The validity of our design decisions and the results may be judged by inspecting the trajectories and ground truth matrices in Figures 6–11.

3.2.2. SLAM Scenario

As a second type of evaluation, we also consider how the method would fare in a SLAM application. In this case, for each scan $\mathcal{S}$ we consider only the most similar corresponding scan $\bar{\mathcal{S}}$ instead of all other scans. The ground truth in this case is a manual labeling of scans as either "overlapping" (meaning that they were acquired at a place that was visited more than once and therefore should be similar to at least one other scan) or "nonoverlapping" (which is to say that they were seen only once). Because the ground truth

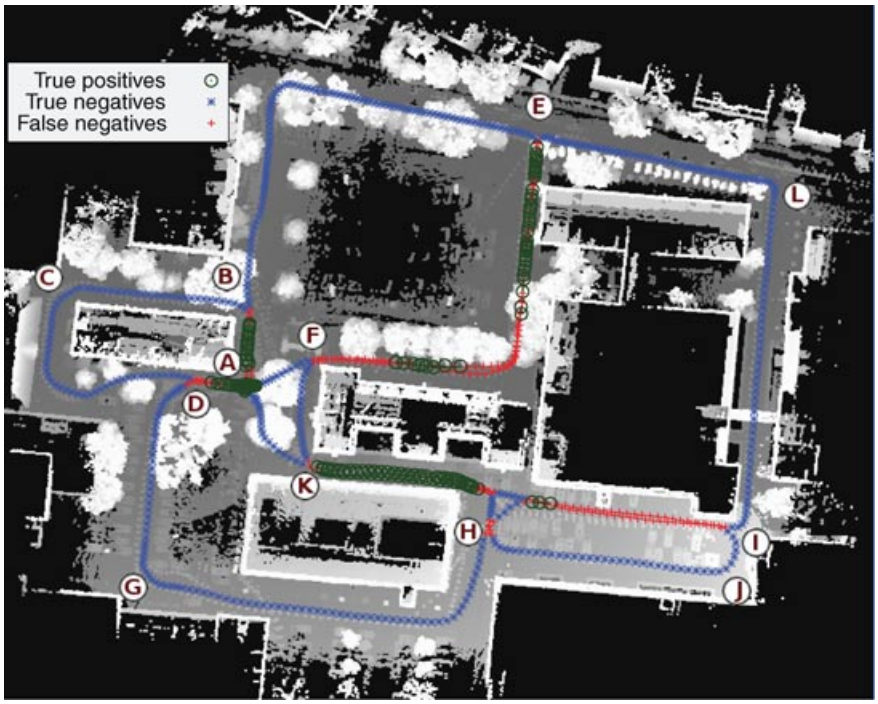
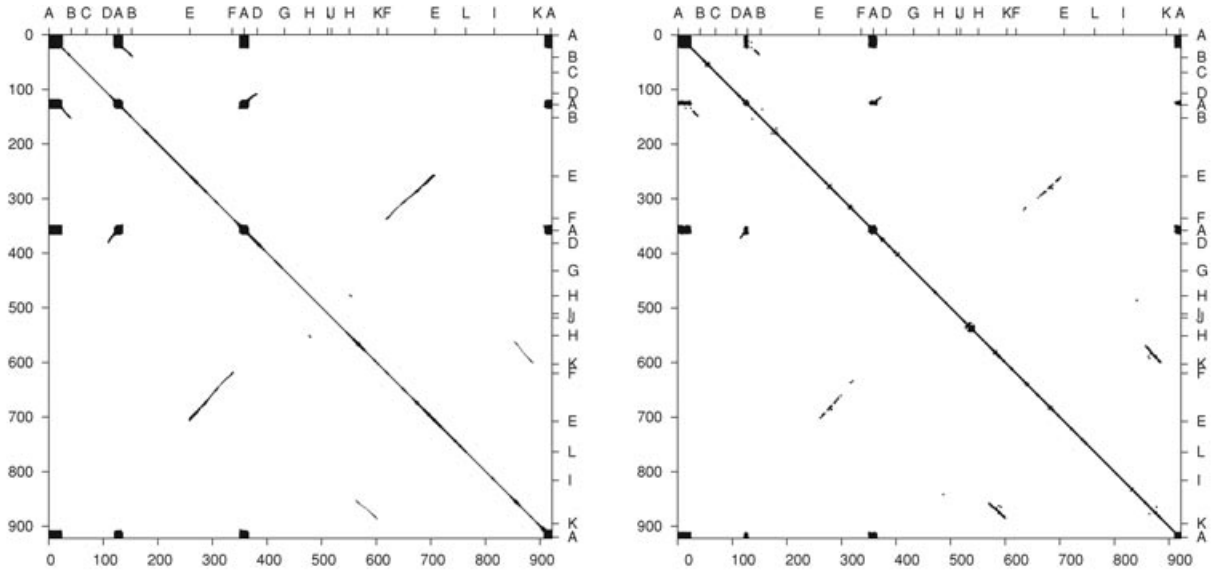


Figure 6. SLAM result for the Hannover2 data set. The robot traveled along the sequence A-B-C-D-A-B-E-F-A-D-G-H-I-J-H-K-F-E-L-I-K-A. Note that there are no false positives and that all true positives are matched to nearby scans.



(a) Thresholded ground truth distance matrix of Hannover2, showing all scan pairs taken less than 3 m apart

(b) Thresholded similarity matrix, showing all scan pairs whose difference value $\tau < 0.0737$

Figure 7. Comparing the ground truth matrix and the output similarity matrix for Hannover2. Scan numbers are on the left-hand and bottom axes; place labels are on the top and right-hand axes. (Because of the large matrix and the small print size, the right-hand image has been morphologically dilated by a 3×3 element in order to better show the values.)

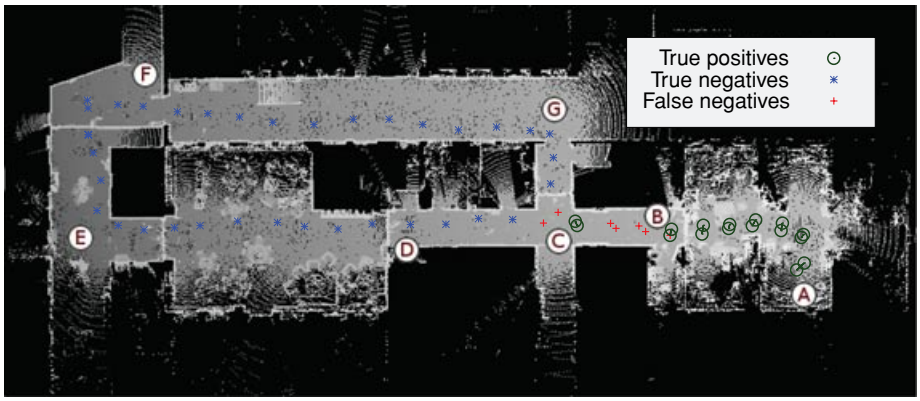
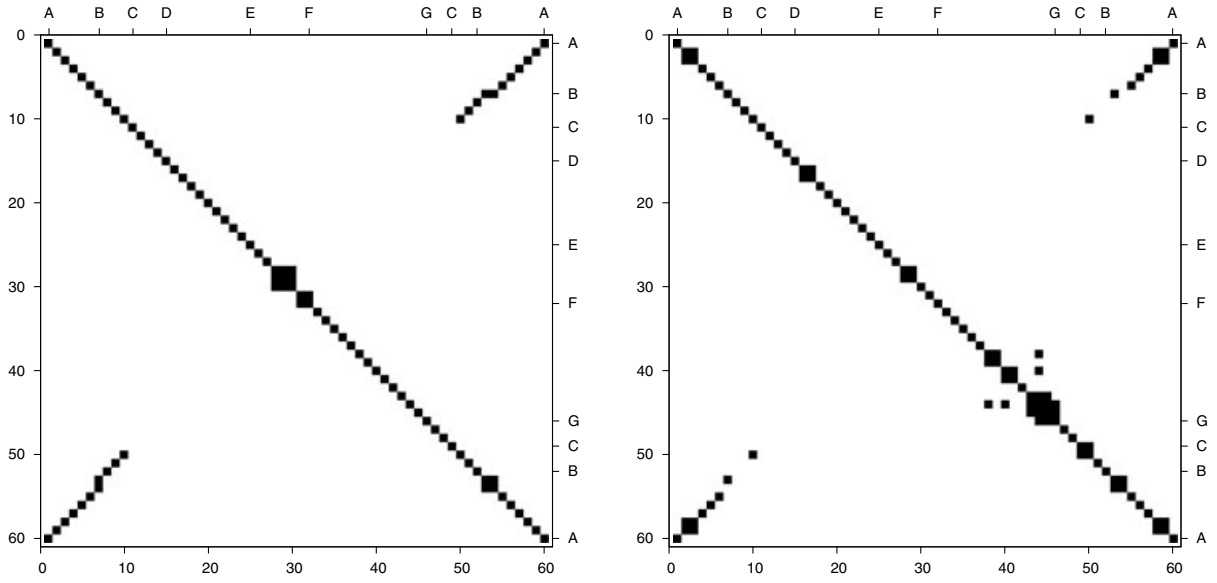


Figure 8. Result for the AASS-loop data set. The robot moved along the path A-B-C-D-E-F-G-C-B-A.

labeling is done to the set of individual scans instead of all combinations of scan pairs, it is feasible to perform manually.

This second type of evaluation is more similar to how the FAB-MAP method of Cummins and Newman (2007, 2008a, 2008b, 2009) has been evaluated. If $\mathcal{S}$ has been labeled as overlapping, the most similar scan $\bar{\mathcal{S}}$ is within 10 m of $\mathcal{S}$ and the difference measure of the two scans is below the threshold $[\tau(\mathcal{S}, \bar{\mathcal{S}}) < t_d]$, then $\mathcal{S}$ is considered a true positive. The 10-m distance threshold is the same as that used by Valgren and Lilienthal (2007) for establishing successful localization. Cummins and Newman (2009) use a 40-m threshold, but that was deemed too



(a) Ground truth distance matrix, showing all scan pairs taken less than1 m apart

(b) Thresholded similarity matrix, showing all scan pairs with a difference value $\tau < 0.099$

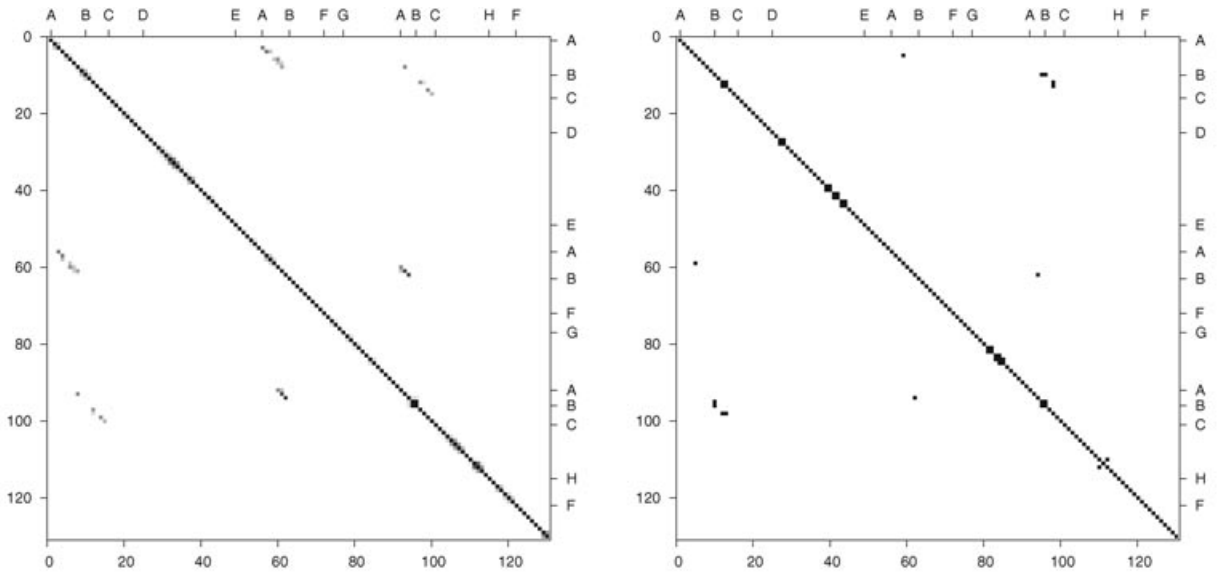
Figure 9. Comparing the ground truth matrix and output similarity matrix for AASS-loop.

Figure 10. Result Kvarntorp. The robot traveled along the sequence A-B-C-D-E-A-B-F-G-A-B-C-H-F-H.

large for the data sets used here. Most of the detected scan pairs are comfortably below the 10-m threshold. For the AASS-loop and Kvarntorp data sets, the maximum interscan distance at detected loop closure is 2.6 m. For Hannover2, 98% of the detected scans are within 5 m of each other, and 83% are within 3 m.

For these experiments, we report the result as precision-recall rates. *Precision* is the ratio of true positive loop detections to the sum of all loop detections. *Recall* is the ratio of true positives to the number of ground truth loop closures. A nonoverlapping scan cannot contribute to the true positive rate, but it can generate a false positive, thus affecting precision. Likewise true loop closures that are incorrectly regarded as negative decrease the recall rate but do not impact the precision rate. It is important to realize that a 1% false-positive rate is not the same as 99% precision. If the number of nonoverlapping scans is much larger than the number of overlapping ones, as is the case for our data sets, falsely detecting 1% of the nonoverlapping ones as positive will decrease the precision rate with much more than 1%.

In a SLAM application, even a single false positive can make the map unusable if no further



(a) Ground truth distance matrix of Kvarntorp, showing all scan pairs taken less than 3 m apart and with an orientation difference of max 20 deg

(b) Thresholded similarity matrix of Kvarntorp, showing all scan pairs with a difference value $\tau < 0.0870$

Figure 11. Comparing the distance matrix and the output similarity matrix for Kvarntorp.

measures are taken to recover from false scan correspondences. Therefore the best difference threshold in this case is the largest possible value with 100% precision.

In this second type of evaluation, we also employed a minimum loop size. Even though we do not use pose estimates from odometry, we assume that the scans are presented as an ordered sequence, successively acquired by the robot as it moves along its trajectory. When finding the most similar correspondence of $\mathcal{S}$, it is compared only to scans that are more than 30 steps away in the sequence. The motivation for this limit is that in the context of a SLAM application it is not interesting to find small "loops" with only consecutive scans. We want to detect loop closure only when the robot has left a place and returned to it later. A side effect of the minimum loop size limit is that some similar scans that are from the same area but more than 10 m apart and therefore otherwise would decrease the precision are removed. However, in a SLAM scenario it makes sense to add such a limit if it is known that the robot cannot possibly close a "real" loop in only a few steps.

Table I. Summary of loop detection results for all scan pairs.

Set	t_r (m)	ol	nol	t_d	Recall (%)
Hannover2	3	9,984	839,178	0.1494	**80.6**
AASS-loop	1	32	3,508	0.0990	**62.5**
Kvarntorp	3	138	16,632	0.1125	**27.5**

This table shows the maximum achievable recall rate with less than 1% false positives and the difference threshold (t_d) at which this is attained. The distance threshold applied to the ground truth matrix is denoted t_r, and the ground truth numbers of overlapping (ol) and nonoverlapping (nol) scan pairs after applying t_r are also shown.

In our previous work on loop detection (Magnusson, Andreasson, et al., 2009) we used this SLAM-type evaluation but obtained the ground truth labeling of overlapping and nonoverlapping scans using a distance threshold. The manual labeling employed here is a better criterion for judging which scans are overlapping and not. Again, refer to the figures visualizing the results (Figures 6, 8, and 10) to judge the validity of the evaluations.

3.3. Results

This section details the results of applying our loop detection method to the data sets described above. The results are summarized in Tables I and II, where the recall rates are shown in boldfaced type.

3.3.1. Hannover2

The Hannover2 data set is the one that is most similar to the kind of outdoor semistructured data investigated in many other papers on robotic loop detection (Bosse & Zlot, 2008b; Cummins & Newman, 2008b; Granström et al., 2009; Valgren & Lilienthal, 2007).

When evaluating the full similarity matrix, the maximum attainable recall rate with at most 1% false positives is 80.6%, using $t_d = 0.1494$. Figure 7(a) shows the ground truth distance matrix of the Hannover2 scans, and Figure 7(b) shows the similarity matrix obtained with the proposed appearance descriptor and difference measure. Note that the two matrices are strikingly similar. Most of the overlapping (dark) parts in the ground truth matrix are captured correctly in the similarity matrix. The distance threshold t_r was set to 3 m.

For the SLAM-style experiment, the maximum recall rate at 100% precision is 47.0%, using $t_d = 0.0737$. The result is visualized in Figure 6, showing

Table II. Summary of SLAM scenario loop detection results.

			Manual threshold				Automatic threshold			
Set	ol	nol	t_d	Recall (%)	Prec. (%)	P(fp) (%)	t_d	Recall (%)	Prec. (%)	P(fp) (%)
Hannover2	428	494	0.0737	**47.0**	100	0.08	0.0843	**55.6**	94.8	0.5
AASS-loop	23	37	0.0990	**69.6**	100	1.17	0.0906	**60.9**	100	0.5
Kvarntorp	35	95	0.0870	**28.6**	100	0.65	0.0851	**22.9**	100	0.5

Precision and recall rates are shown both for manually selected t_d and for thresholds selected using a gamma mixture model, as described in Section 3.4. The probability of false positives P(fp) according to the mixture model is shown for both thresholds. The numbers of (ground truth) overlapping and nonoverlapping scans for each set are denoted ol and nol.

all detected true positives and the scans that they are matched to, as well as true and false negatives.

If no minimum loop size is used in the SLAM evaluation (thus requiring that the robot should be able to relocalize itself from the previous scan at all times), the maximum recall rate at 100% precision is 24.6% at $t_d = 0.0579$. If the same difference threshold as above is used ($t_d = 0.0737$), the recall rate for this case is 45.7% and the precision rate is 98.6%, with six false correspondences (0.65% of the 922 scans). Of the six errors, four scans (two pairs) are from the parking lot between locations H and J, which is a place with repetitive geometric structure. The other two are from two corners of the same building: locations A and B.

At this point it should be noted that even a recall rate of around 30% often is sufficient to close all loops in a SLAM scenario, as long as the detected loop closures are uniformly distributed over the trajectory, because several scans are usually taken from each location. Even if one overlapping scan pair is not detected (because of noisy scans, discretization artifacts in the surface shape histograms, or dynamic changes), one of the next few scans is likely to be detected instead. [This fact has also been noted by Cummins and Newman (2008b) and Bosse and Zlot (2008b).]

As a side note, we would also like to mention that using scan registration alone to detect loop closure is not sufficient for this data set, as was described by Wulf et al. (2007). Because they depend on an accurate initial pose estimate (which is necessary even for reliable and fast scan registration algorithms), it is necessary to use the robot's current pose estimate and consider only the closest few scans to detect loop closure. Therefore the method of Wulf et al. (2007), and indeed all methods using local pairwise registration methods such as ICP or 3D-NDT, cannot detect loops when the accumulated pose error is too large. In contrast, the method proposed in this text requires no pose information.

3.3.2. AASS-Loop

When evaluating the full similarity matrix for the AASS-loop data set, we used a distance threshold t_r of 1 m instead of 3 m on the ground truth distance matrix. The reason for the tighter distance threshold in this case is the many passages and tight corners of this data set. The appearance of scans often changes drastically from one scan to the next when rounding a corner into another corridor or passing through a door, and an appearance-based loop detection method cannot be expected to handle such scene changes. The 1-m threshold filters out all such scan pairs while keeping the truly overlapping scan pairs that occur after the robot has returned to location C, as can be seen in Figure 9(a).

For this data set, the maximum recall rate (for the complete similarity matrix) with less than 1% false positives is 62.5% ($t_d = 0.0990$). In the SLAM scenario, the recall rate for this data set was 69.6% at 100% precision, using $t_d = 0.099$.

The part of this data set that contains a loop closure (between locations A and C) is traversed in the opposite direction when the robot returns. The high recall rate illustrates that the surface shape histograms are robust to changes in rotation.

The trajectory of the AASS-loop data set is shown in Figure 8. The ground truth and similarity matrices are shown in Figure 9.

3.3.3. Kvarntorp

The Kvarntorp data set had to be evaluated slightly differently, because an omnidirectional scanner was not used to record this data set. An appearance-based loop detection algorithm cannot be rotation invariant if the input scans are not omnidirectional. When looking in opposite directions from the same place, the view is generally very different. Therefore only scans taken in similar directions (within 20 deg) were counted as overlapping when evaluating the algorithm for Kvarntorp. The scans that were taken at overlapping positions but with different orientations were all (correctly) marked as nonoverlapping by the algorithm. With the exception of the way of determining which scans are from overlapping sections, the same evaluation and algorithm parameters were used for this data set as for Hannover2.

Evaluating the full similarity matrix, the recall rate at 1% false positives is 27.5% ($t_d = 0.1134$). For the SLAM experiment, $t_d = 0.0870$ gives the highest recall rate at full precision: 28.6%.

The challenging properties of the underground mine environment are shown in the substantially lower recall rates for this data set compared to Hannover2. Still, a reasonable distribution of the overlapping scans in the central tunnel is detected in the SLAM scenario (shown in Figure 10), and there are no false positives. The ground truth distance matrix is shown in Figure 11(a), and the similarity matrix is shown in Figure 11(b). Comparing the two figures,

it can be seen that some scans are recognized from all overlapping segments: For all off-diagonal stripes in Figure 11(a), there is at least one corresponding scan pair below the difference threshold in Figure 11(b).

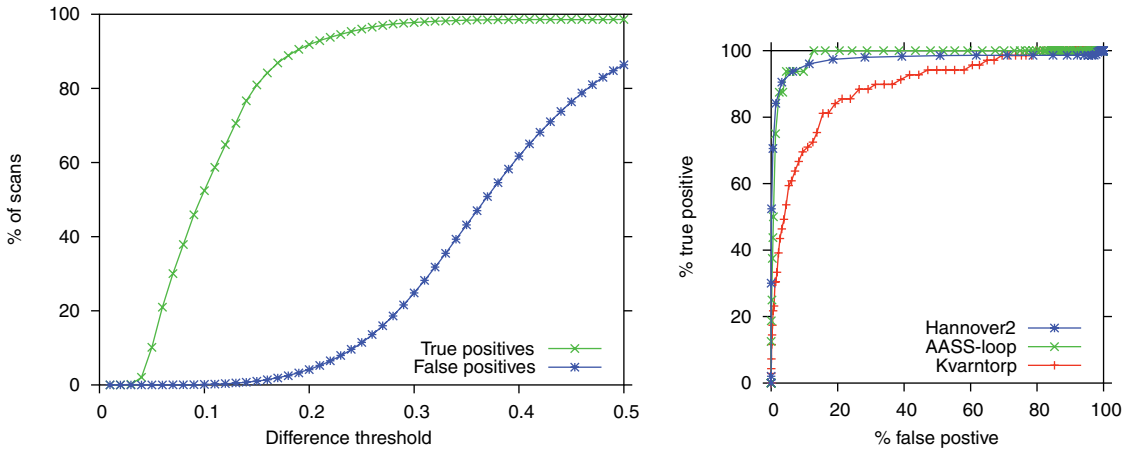
3.4. Automatic Threshold Selection

It is important to find a good value for the difference threshold t_d. Using a too-small value results in a small number of true positives (correctly detected overlapping scans). Using a too-large value results in false positives (scan pairs considered overlapping even though they are not). Figures 12 and 13 illustrate the discriminative ability of the surface shape histograms for the two different modes of evaluation, showing how the numbers of true positives and errors change with increasing values of the difference threshold, as well as the ROC (receiver operating characteristics) curve.

The results reported thus far used manually chosen difference thresholds, selected with the help of the available ground truth data. To determine t_d when ground truth data are unavailable, it is desirable to estimate the distributions of difference values [Eq. (13)] for overlapping scans versus the values for nonoverlapping scans. Given the set of numbers containing all scans' smallest difference values, it can be assumed that the values are drawn from two distributions—one for the overlapping scans and one for the nonoverlapping ones. If it is possible to fit a probabilistic mixture model of the two components to the set of values, a good value for the difference threshold should be such that the estimated probability of false positives $P(\text{fp})$ is small but the estimated probability of true positives is as large as possible. Figure 14 shows a histogram of the difference values for the scans in the Hannover2 data set. The histogram was created using the difference value of most similar scan for each scan in the data set. (In other words, this is the outcome of the algorithm in the SLAM scenario.) The figure also shows histograms for the overlapping and nonoverlapping subsets of the data (which are not known in advance).

A common way to estimate mixture model parameters is to fit a Gaussian mixture model to the data with the expectation maximization (EM) algorithm. However, inspecting the histograms of difference values (as in Figure 14), it seems that the underlying distributions are not normally distributed but have a significant skew, with the right tail being longer than the left. As a matter of fact, trying to fit a two-component Gaussian mixture model with EM usually results in distribution estimates with too-large means. It is sometimes feasible to use three Gaussian components instead, where one component is used to model the long tail of the skew data (Magnusson, Andreasson, et al., 2009). However, only a binary classification is desired, so there is no theoretical ground for such a model.

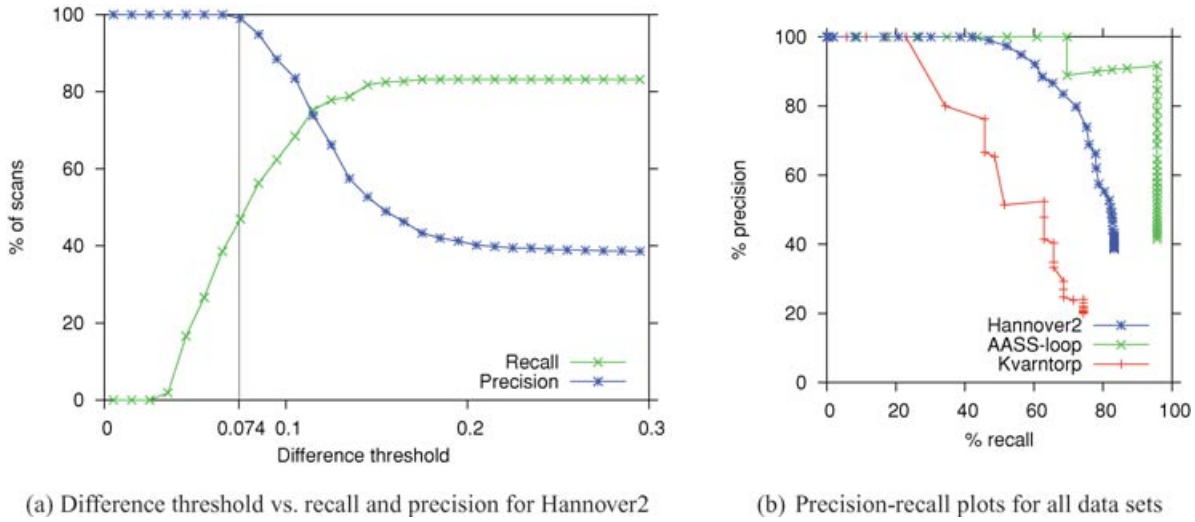
Gamma distributed components fit the difference value distributions better than Gaussians. Figure 15(a) shows two gamma distributions fitted in



(a) Difference threshold vs. true and false positive rates for Hannover2

(b) ROC plots for all data sets

Figure 12. Plots of the appearance descriptor's discriminative ability, evaluating all possible scan pairs.



(a) Difference threshold vs. recall and precision for Hannover2 (b) Precision-recall plots for all data sets

Figure 13. Plots of the appearance descriptor's discriminative ability for the SLAM scenario. In (a), the best threshold (giving the maximum number of true positives at 100% precision) is marked with a bar.

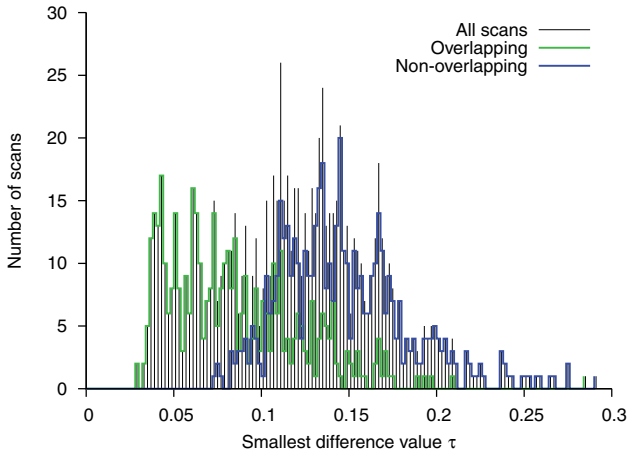


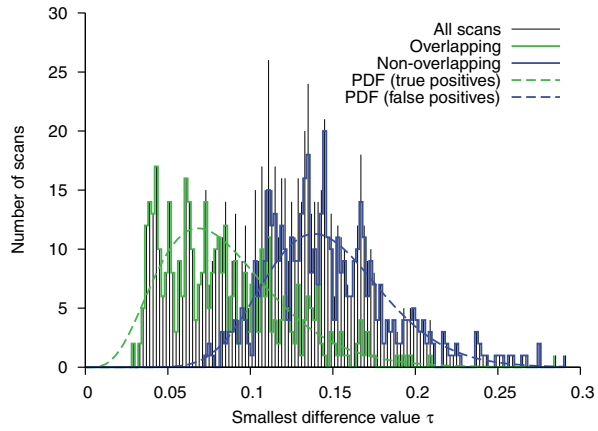
Figure 14. Histograms of the smallest difference values for overlapping and nonoverlapping scans of the Hannover2 data set. (In general, only the histogram for all scans is known.)

isolation to each of the two underlying distributions. Because the goal is to choose t_d such that the expected number of false positives is small, a reasonable criterion is that the cumulative distribution function of the mixture model component that corresponds to nonoverlapping scans should be small. This is equivalent to saying that $P(\text{fp})$ should be small. Figure 15(b) shows the cumulative distribution functions of the mixture model components in Figure 15(a).

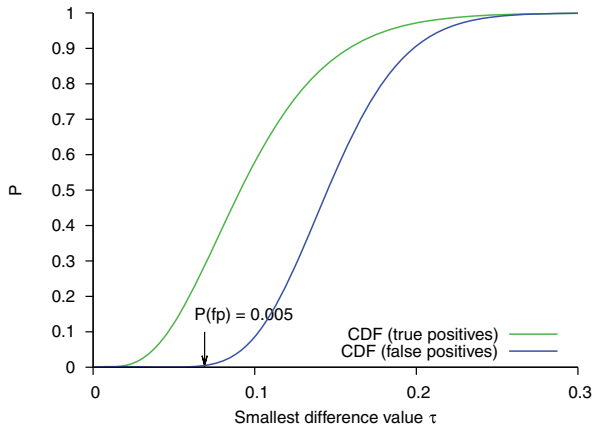
For the Kvarntorp data set, EM finds a rather well-fitting mixture model. With $P(\text{fp}) = 0.005$ the threshold value is 0.0851, resulting in a 22.9% recall rate with no false positives. This is a slightly conservative threshold, but it has 100% precision.

The AASS-loop data set is more challenging for EM. It contains only 60 scans, which makes it difficult to fit a reliable probability distribution model to the difference values. Instead, the following approach was used for evaluating the automatic threshold selection. Two maximum likelihood gamma distributions were fitted to the overlapping and nonoverlapping scans separately. Using these distributions and the relative numbers of overlapping and nonoverlapping scans of AASS-loop, 600 gamma distributed random numbers were generated, and EM was applied to find a maximum likelihood model of the simulated combined data. The simulated values represent the expected output of collecting scans at a much denser rate in the same environment. Using the resulting mixture model and $P(\text{fp}) = 0.005$ gave $t_d = 0.091$ and a recall rate of 60.9% using the AASS-loop scans.

Because EM is a local optimization algorithm, it can be sensitive to the initial estimates given. When applied to the output of the Hannover2 data set, it tends to converge to one of two solutions, shown in



(a) Two maximum likelihood gamma distributions fitted to the overlapping and nonoverlapping scans in isolation

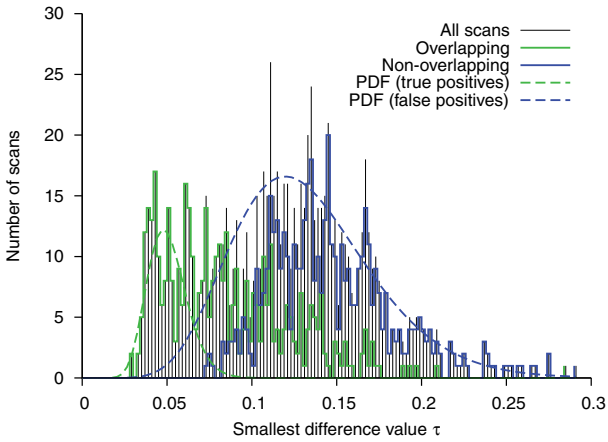


(b) Cumulative distribution functions of the mixture model components, showing the threshold such that the estimated probability of false positives is 0.5%

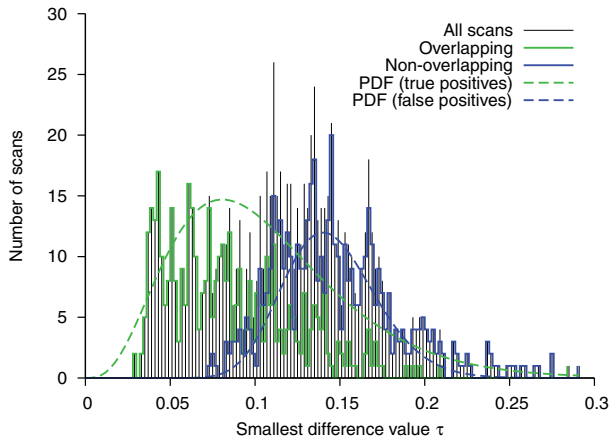
Figure 15. Determining t_d for the Hannover2 data set using a gamma mixture model.

Figure 16. From visual inspection, the solution of Figure 16(b) looks better than that of Figure 16(a), but the likelihood function of the solution in Figure 16(a) is higher. Solution 16(a) uses a wider than necessary model of the nonoverlapping scans, resulting in a conservative threshold value. With $P(\text{fp}) = 0.005$, solution 16(a) gives $t_d = 0.0500$ and only a 20.8% recall rate, although at 100% precision. The numbers for solution 16(b) are $t_d = 0.0843$, 55.6% recall, and 94.8% precision. Table II includes the results of solution 16(b).

This approach for determining t_d involves no training and is a completely unsupervised learning process. However, the difference threshold can only be estimated offline: not because of the computational burden (which is very modest) but because a



(a) One solution



(b) Another solution

Figure 16. Histograms of the difference values τ (considering each scan's most similar correspondence) for overlapping and nonoverlapping scans of the Hannover2 data set. The components of a gamma mixture model fitted with EM for two different initial parameter estimates are also shown. The log-likelihood ratio of 16(a)/16(b) is 1.01.

sufficiently large sample of scans must have been encountered before EM can be used to estimate a reliable threshold. As long as there are enough samples, the method described in this section gives a useful estimate for t_d. However, because it is not possible to guarantee that the output threshold value produces no false positives, a reliable SLAM implementation should still have some way of handling spurious false positives.

3.5. Execution Time

The experiments were run using a C++ implementation on a laptop computer with a 1,600-MHz Intel Celeron CPU and 2 GB of RAM.

For the AASS-loop data set, average times (measured with the gprof profiling utility) for computing the surface shape histograms were 0.5 s per call to the histogram computation function and in total 2.2 s per scan to generate histograms (including transforming the point cloud, generating $\boldsymbol{f}'$ and the histograms that make up $\mathcal{F}$). The average number of histograms required for rotation invariance (that is, the size of $\mathcal{F}$) is 2.4. In total, 0.14 s was spent computing similarity measures for scan pairs. There are 60 scans in the data set, 144 histograms were created, and $144^2 = 20{,}736$ similarity measures were computed, so the average time per similarity comparison [Eq. (11)] was around 7 μs, and it took less than 0.5 ms to compare two scans [Eq. (13)]. (Naturally, we could also have computed only one-half of the similarity matrix, because the matrix is symmetric.) In other words, once the histograms have been created, if each scan requires the generation of 2.4 histograms on average, a new scan can be compared to roughly 25,000 other scans in 1 s when testing for loop closure, using exhaustive search. The corresponding numbers for all of the data sets are shown in Table III.

The time for creating the histograms and the number of histograms required for rotation invariance depend on the data, but the time required for similarity comparisons is independent of the data.

The time spent on histogram creation can be significantly reduced if transformations are applied to the first computed histogram when creating $\mathcal{F}$, instead of computing new histograms from scratch after transforming the original point cloud. With this optimization, the total time spent while generating the appearance descriptor is 1.0 s per scan instead of 2.2 s per scan for the AASS-loop data set. However, the resulting histograms are not identical to the ones that are achieved by recomputing histograms from the transformed point clouds. They are only approximations. For all three data sets, the recognition results were marginally worse when using this optimization.

4. RELATED WORK

4.1. Other Loop Detection Approaches

A large part of the related loop detection literature is focused on data from camera images and 2D range data.

Ramos, Nieto, and Durrant-Whyte (2007) used a combination of visual cues and laser readings to associate features based on both position and appearance. They demonstrated that their method works well in outdoor environments with isolated features. The experiments used for validation were performed on data collected in Victoria Park, Sydney, where the available features are sparsely planted trees. A limitation of the method of Ramos et al. is that the laser features are found by clustering isolated point segments, which are stored as curve segments. In many other settings (such as indoor or urban environments), the appearance of scans is quite different from the ones in Victoria Park in that features are not generally surrounded by empty space. Compared to the laser features used by Ramos et al., the proposed surface

Table III. Summary of resource requirements.

Data set	Scans	Points/scan	Avg. histogram creation time (s)	Avg. histograms/scan
Hannover2	922	15,000	0.18	3.2
Kvarntorp	130	70,000	0.27	2.8
AASS-loop	60	112,000	0.50	2.4

In addition to the number of scans in each data set and the average point count per scan, the table shows the average time to create a single histogram (on a 1.6-GHz CPU) and the average number of histograms per scan.

shape histograms have the advantage that they require no clustering of the input data and therefore it is likely that they are more context independent. It is currently not clear how the method of Ramos et al. would perform in a more cluttered environment.

Cummins and Newman (2007, 2008a, 2008b, 2009) have published several articles on visual loop detection using their FAB-MAP method. They use a bag-of-words approach in which scenes are represented as a collection of "visual words" (local visual features) drawn from a "dictionary" of available features. Their appearance descriptor is a binary vector indicating the presence or absence of all words in the dictionary. The appearance descriptor is used within a probabilistic framework together with a generative model that describes how informative each visual word is by the common co-occurrences of words. In addition to simple matching of appearance descriptors, as has been done in the present work, they also use pairwise feature statistics and sequences of views to address the perceptual aliasing problem. Cummins and Newman (2008) have reported recall rates of 37%–48% at 100% precision, using camera images from urban outdoor data sets. Recently Cummins and Newman (2009) reported on the experiences of applying FAB-MAP on a very large scale, showing that the computation time scales well to trajectories as long as 1,000 km. The precision, however, is much lower on the large data set, as is to be expected.

A method that is more similar to the approach presented here is the 2D histogram matching of Bosse and Roberts (2007) and Bosse and Zlot (2008a, 2008b). Although our loop detection method may also be referred to as histogram matching, there are several differences. For example, Bosse et al. use the normals of oriented points instead of the orientation/shape features of NDT. Another difference lies in the amount of discretization. Bosse et al. create 2D histograms with one dimension for the spatial distance to the scan points and one dimension for scan orientations. The angular histogram bins cover all possible rotations of a scan in order to achieve rotation invariance. Using 3-deg angular resolution and 1-m range resolution, as in the published papers, results in $120 \times 200 = 240{,}000$ histogram bins for the 2D case. For unconstrained 3D motion with angular bins for the x, y, and z axes, a similar discretization would lead to many millions of bins. In contrast, the 3D histograms presented here require only a few dozen bins. At a false-positive rate of 1%, Bosse and Zlot (2008b) have achieved a recall rate of 51% for large urban data sets, using a manually chosen threshold.

Very recent work by Granström et al. (2009) showed good performance of another 2D loop detection algorithm. Their method uses AdaBoost (Freund & Schapire, 1997) to create a strong classifier composed from 20 weak classifiers, each of which describes a global feature of a 2D laser scan. The two most important weak classifiers are reported to be the area enclosed by the complete 2D scan and the area where the scan points with maximum range have been removed. With 800 scan pairs manually selected from larger urban data sets (400 overlapping pairs and 400 nonoverlapping ones), Granström et al. report an 85% recall rate with 1% false positives. It would be interesting to see how their method could be extended to the 3D case and how it would perform in other environments.

Perhaps the most relevant related method for loop detection from 3D range data is the work by Johnson (1997) and Huber (2002). Johnson's "spin images" are local 3D feature descriptors that give detailed descriptions of the local surface shape around an oriented point. Huber (2002) has described a method based on spin images for matching multiple 3D scans without initial pose estimates. Such global registration is closely related to the loop detection problem. The initial step of Huber's multiview surface matching method is to compute a model graph by using pairwise global registration with spin images for all scan pairs. The model graph contains potential matches between pairs of scans, some of which may be incorrect. Surface consistency constraints on sequences of matches are used to reliably distinguish correct matches from incorrect ones because it is not possible to distinguish the correct and incorrect matches at the pairwise level. Huber has used this method to automatically build models of various types of scenes. However, we are not aware of a performance measurement that is comparable to the work covered in this paper. Our algorithm can be seen as another way of generating the initial model graph and evaluating a local quality measure. An important difference between spin images and the surface shape histograms proposed here is that spin images are local feature descriptors, more akin to visual words, describing the surface shape around one point. In contrast, the surface shape histograms are global appearance descriptors, describing the appearance of a whole 3D point cloud. Comparing spin images to the local Gaussian features used in this work,

spin images are more descriptive and invariant to rotation when reliable point normals are available. Normal distributions are unimodal functions, whereas spin images can capture arbitrary surface shapes if the resolution is high enough. However, the processing requirements are quite different for the two methods. Using data sets containing 32 scans with 1,000 mesh faces each, as done by Huber (2002), the time to compute the initial model graph using spin-image matching can be estimated to $1.5 \times 32^2 = 1,536$ s. (The complete time is not explicitly stated, but pairwise spin-image matching is reported to require 1.5 s on average.) With a data set of that size, a rough estimate of the execution time of the algorithm proposed in this paper is $32 \times 0.8 + (32 \times 3)^2 \times 7 \times 10^{-6} = 26$ s on similar hardware, based on the execution times in Table III. On a data set of a more realistic size, the difference would be even greater.

4.2. Comparing Results

As discussed in Section 3.2., it is not always obvious how to determine ground truth in the context of loop detection. Granström et al.(2009) solved this problem by evaluating their algorithm on a selection of 400 scan pairs that were manually determined to be overlapping and 400 nonoverlapping ones. However, we would like to evaluate the performance on the complete data sets. Bosse and Roberts (2007) and Bosse and Zlot (2008a, 2008b) use the connectivity graph between submaps created by the Atlas SLAM framework (Bosse, Newman, Leonard, & Teller, 2004) as the ground truth. In this case, each scan has a single correspondence in each local subsequence of scans (although there may be other correspondences at subsequent revisits to the same location). In our evaluations of the full similarity matrix for each data set no such preprocessing was performed. Instead, we applied a narrow distance threshold t_r to the scan-to-scan distance matrix in order to generate a ground truth labeling of true and false positives. The fact that the approaches used to determine the ground truth vary so much between different authors makes it difficult to compare the results.

Furthermore, because all of the methods discussed above were evaluated on different data sets, it is not possible to make any conclusive statements about how the quality of the results compare to one another, both because the appearances of scans may vary greatly between different data sets and also because the relative numbers of overlapping and nonoverlapping scans differ. A false-positive rate of 1% (of all nonoverlapping scans) for a data set that has a large ratio of nonoverlapping scans is not directly comparable to the same result for a set with more loop closures.

Having said that, we will still compare our results to those reported in the related literature in order to give some indication of the relative performance of our approach. On the Hannover2 data set, which is the only one with characteristics comparable to those used in the related work, we obtained a recall rate of 80.5% at 1% false positives when evaluating all scan pairs. This result compares well to the 51% recall rate of Bosse and Zlot (2008b) and the 85% recall rate of Granström et al.(2009).

The SLAM-style experiment on the same data set is more similar to those of Cummins and Newman (2007, 2008a, 2008b). With no false positives, we achieved a 47.0% recall rate for the Hannover2 data set, which is comparable to their recall rates of 37%–48%.

5. SUMMARY AND CONCLUSIONS

We have described a new approach to appearance-based loop detection from 3D range data by comparing surface shape histograms. Compared to 2D laser-based approaches, using 3D data makes it possible to avoid dependence on a flat ground surface. However, 3D scans bring new problems in the form of a massive increase in the amount of data and more complicated rotations, which means a much larger pose space in which to compare appearances. We have shown that the proposed surface shape histograms overcome these problems by allowing for drastic compression of the input 3D point clouds while being invariant to rotation. We propose to use EM to fit a gamma mixture model to the output similarity measures in order to automatically determine the threshold that separates scans at loop closures from nonoverlapping ones, and we have shown that doing so gives threshold values that are in the vicinity of the manually selected ones that give the best results for all our data sets. With experimental evidence we have shown that the presented approach can achieve high recall rates at low false-positive rates in different and challenging environments. Another contribution of this work is that we have focused on the problem of providing quantifiable performance evaluations in the context of loop detection. We have discussed the difficulties of determining unambiguous ground truth

correspondences that can be compared for different loop detection approaches.

We can conclude that NDT is a suitable representation for 3D range scans. It allows for very good scan registration (Magnusson et al., 2007; Magnusson, Nüchter, et al., 2009) and can be used as an intermediate representation for the proposed surface shape histograms. We also conclude that the proposed NDT-based surface shape histograms perform well in comparison with related loop detection methods based on 2D and 3D range data as well as current methods using visual data. The highly compact histogram representation (which uses 50–200 values on average to represent a 3D point cloud with several tens of thousands of points) makes it possible to compare scans very quickly. We can compare a 3D scan to around 25,000 others in 1 s, as compared to 1.5 s per comparison using 3D spin-image descriptors. The high speed makes it possible to detect loop closures even in large maps by exhaustive search, which is an important contribution of our work. Even though the input data are highly compressed, the recall rate is still 80.5% at a 1% false-positive rate for our outdoor campus data set.

6. FUTURE WORK

It would be very interesting in future work to compare our approach to different methods using the same data set. The Kvarntorp data set includes 2D scans and camera images in addition to the 3D scans used here, so that data set would lend itself especially well to comparing different approaches.

It would be equally important to improve current experimental methodology to include a unified method for selecting true and false positives in the context of loop detection. A formal definition of what constitutes "a place" in this context would be very welcome, for the same purpose.

To further improve the performance of our approach, future work could include learning a generative model in order to learn how to disregard common nondiscriminative features (such as floor and ceiling orientations), based on the general appearance of the current surroundings, as done previously in the visual domain (Cummins & Newman, 2008b). The disjoint range intervals that we have used in the present work may be a source of error, although the use of overlapping NDT cells should alleviate also these discretization artifacts to some degree. We will consider overlapping range intervals in future work. It would also be interesting to do a more elaborate analysis of the similarity matrix than applying a simple threshold, in order to better discriminate between overlapping and nonoverlapping scans and to evaluate the effects of difference measures other than the one used here. Another potential direction is to research whether it is possible to learn more of the parameters from the data. Further future work should include investigating how the proposed loop detection method is affected by dynamic changes, such as moving vehicles or people.

REFERENCES

Biber, P., & Straßer, W. (2003, October). The normal distributions transform: A new approach to laser scan matching. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV (pp. 2743–2748).

Booij, O., Terwijn, B., Zivkovic, Z., & Kröse, B. (2007, April). Navigation using an appearance based topological map. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Rome, Italy (pp. 3927–3932). IEEE.

Borrmann, D., Elseberg, J., Lingemann, K., Nüchter, A., & Hertzberg, J. (2008). Globally consistent 3D mapping with scan matching. Journal of Robotics and Autonomous Systems, 56(2), 130–142.

Bosse, M., Newman, P., Leonard, J., & Teller, S. (2004). Simultaneous localization and map building in large-scale cyclic environments using the Atlas framework. International Journal of Robotics Research, 23(12), 1113–1139.

Bosse, M., & Roberts, J. (2007, April). Histogram matching and global initialization for laser-only SLAM in large unstructured environments. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Rome, Italy (pp. 4820–4826).

Bosse, M., & Zlot, R. (2008a, June). Keypoint design and evaluation for global localization in 2D LIDAR maps. In Robotics: Science and Systems, Zurich, Switzerland.

Bosse, M., & Zlot, R. (2008b). Map matching and data association for large-scale two-dimensional laser scan-based SLAM. International Journal of Robotics Research, 27(6), 667–691.

Cummins, M., & Newman, P. (2007, April). Probabilistic appearance based navigation and loop closing. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Rome, Italy (pp. 2042–2048).

Cummins, M., & Newman, P. (2008a, May). Accelerated appearance-only SLAM. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Pasadena, CA (pp. 1828–1833).

Cummins, M., & Newman, P. (2008b). FAB-MAP: Probabilistic localization and mapping in the space of

appearance. International Journal of Robotics Research, 27(6), 647–665.

Cummins, M., & Newman, P. (2009, June). Highly scalable appearance-only SLAM—FAB-MAP 2.0. In Robotics: Science and Systems, Seattle, WA.

Frese, U., Larsson, P., & Duckett, T. (2005). A multilevel relaxation algorithm for simultaneous localisation and mapping. IEEE Transactions on Robotics, 21(2), 196–207.

Frese, U., & Schröder, L. (2006, October). Closing a million-landmarks loop. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), Beijing, China (pp. 5032–5039).

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139.

Granström, K., Callmer, J., Ramos, F., & Nieto, J. (2009, May). Learning to detect loop closure from range data. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan (pp. 15–22).

Grisetti, G., Grzonka, S., Stachniss, C., Pfaff, P., & Burgard, W. (2007, October). Efficient estimation of accurate maximum likelihood maps in 3D. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), San Diego, CA (pp. 3472–3478).

Huber, D. F. (2002). Automatic three-dimensional modeling from reality. Ph.D. thesis, Carnegie Mellon University.

Johnson, A. E. (1997). Spin images: A representation for 3-D surface matching. Ph.D. thesis, Carnegie Mellon University.

Konolige, K., Bowman, J., Chen, J. D., Mihelich, P., Calonder, M., Lepetit, V., & Fua, P. (2009, June). View-based maps. In Robotics: Science and Systems, Seattle, WA.

Magnusson, M., Andreasson, H., Nüchter, A., & Lilienthal, A. J. (2009, May). Appearance-based loop detection from 3D laser data using the normal distributions transform. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan (pp. 23–28).

Magnusson, M., Lilienthal, A. J., & Duckett, T. (2007). Scan registration for autonomous mining vehicles using 3D-NDT. Journal of Field Robotics, 24(10), 803–827.

Magnusson, M., Nüchter, A., Lörken, C., Lilienthal, A. J., & Hertzberg, J. (2009, May). Evaluation of 3D registration reliability and speed—A comparison of ICP and NDT. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan (pp. 3907–3912).

Ramos, F. T., Nieto, J., & Durrant-Whyte, H. F. (2007, April). Recognising and modelling landmarks to close loops in outdoor SLAM. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Rome, Italy (pp. 2036–2041).

Saff, E. B., & Kuijlaars, A. B. J. (1997). Distributing many points on a sphere. Mathematical Intelligencer, 19(1), 5–11.

Valgren, C., & Lilienthal, A. J. (2007, September). SIFT, SURF and seasons: Long-term outdoor localization using local features. In Proceedings of the European Conference on Mobile Robots (ECMR), Freiburg, Germany (pp. 253–258).

Wulf, O., Nüchter, A., Hertzberg, J., & Wagner, B. (2007, October). Ground truth evaluation of large urban 6D SLAM. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), San Diego, CA (pp. 650–657).