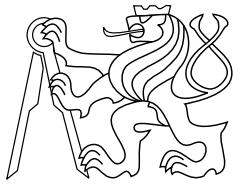




CENTER FOR
MACHINE PERCEPTION



CZECH TECHNICAL
UNIVERSITY IN PRAGUE

BACHELOR THESIS

ISSN 1213-2365

3D Point Cloud Registration, Experimental Comparison and Fusing Range and Visual Data

Aleš Hrabalík

hrabaale@fel.cvut.cz, svoboda@cmp.felk.cvut.cz

May 23, 2014

Thesis Advisor: Tomáš Svoboda

The work was supported by EC project FP7-ICT-609763 TRADR and by the CTU project SGS13/142/OHK3/2T/13. Any opinions expressed in this paper do not necessarily reflect the views of the European Community. The Community is not liable for any use that may be made of the information contained herein.

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: <http://cmp.felk.cvut.cz>

3D Point Cloud Registration, Experimental Comparison and Fusing Range and Visual Data

Aleš Hrabalík

May 23, 2014

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student: Aleš Hrabalík

Studijní program: Otevřená informatika (bakalářský)

Obor: Informatika a počítačové vědy

Název tématu: Prostorová registrace oblaku 3D bodů, experimentální porovnání a využití obrazové informace

Pokyny pro vypracování:

Zopakujte registrační experimenty na ETHZ datasetu [3] s metodou Normal Distribution Transform (NDT) [1], volitelně s navazující metodou Iterative Closest Point (ICP) [2]. Pro metodu NDT jsou k dispozici kódy, dataset i protokol. Implementujte vhodné rozhraní pro napojení různých datasetů a metod. Porovnejte dosažené výsledky s ostatními metodami [3] a [4]. Identifikujte meze funkčnosti algoritmů. Navrhněte vhodné použití obrazové informace za předpokladu známé vzájemné kalibrace kamery a hloubkového senzoru.

Seznam odborné literatury:

- [1] Stoyanov, T.; Magnusson, M. & Lilienthal, A. (2012), Point set registration through minimization of the L2 distance between 3D-NDT models, in 'Robotics and Automation (ICRA), 2012 IEEE International Conference on', pp. 5196-5201.
- [2] Pomerleau, F.; Colas, F.; Siegwart, R. & Magnenat, S. (2013), 'Comparing ICP variants on real-world data sets', Autonomous Robots 34(3), 133-148.
- [3] Pomerleau, F.; Liu, M.; Colas, F. & Siegwart, R. (2012), 'Challenging Data Sets for Point Cloud Registration Algorithms', The International Journal of Robotics Research.
- [4] Petricek, T. & Svoboda, T.: Point Cloud Registration from Local Feature Correspondences-Evaluation on Challenging Datasets, 2014 (unpublished work, under review).

Vedoucí bakalářské práce: doc. Ing. Tomáš Svoboda, Ph.D.

Platnost zadání: do konce letního semestru 2014/2015

L.S.

doc. Dr. Ing. Jan Kybic
vedoucí katedry

prof. Ing. Pavel Ripka, CSc.
děkan

V Praze dne 10. 1. 2014

BACHELOR PROJECT ASSIGNMENT

Student: Aleš Hrabalík

Study programme: Open Informatics

Specialisation: Computer and Information Science

Title of Bachelor Project: 3D Point Cloud Registration, Experimental Comparison and Fusing Range and Visual Data

Guidelines:

Replicate the registration experiments [3] with the Normal Distribution Transform (NDT) method [1]. If proved applicable, use the Iterative Closest Point method [2] for refining the results. Code, dataset and protocol are available for the NDT method. Implement an interface for interconnecting various algorithms and dataset. Compare the achieved results with methods [3,4]. Identify the limits of the tested algorithms. Propose a method that would make use of visual information. Assume known calibration between the imagery and range sensor.

Bibliography/Sources:

- [1] Stoyanov, T.; Magnusson, M. & Lilienthal, A. (2012), Point set registration through minimization of the L2 distance between 3D-NDT models, in 'Robotics and Automation (ICRA), 2012 IEEE International Conference on', pp. 5196-5201.
- [2] Pomerleau, F.; Colas, F.; Siegwart, R. & Magnenat, S. (2013), 'Comparing ICP variants on real-world data sets', Autonomous Robots 34(3), 133-148.
- [3] Pomerleau, F.; Liu, M.; Colas, F. & Siegwart, R. (2012), 'Challenging Data Sets for Point Cloud Registration Algorithms', The International Journal of Robotics Research.
- [4] Petricek, T. & Svoboda, T.: Point Cloud Registration from Local Feature Correspondences- Evaluation on Challenging Datasets, 2014 (unpublished work, under review).

Bachelor Project Supervisor: doc. Ing. Tomáš Svoboda, Ph.D.

Valid until: the end of the summer semester of academic year 2014/2015

L.S.

doc. Dr. Ing. Jan Kybic
Head of Department

prof. Ing. Pavel Ripka, CSc.
Dean

Prague, January 10, 2014

Prohlášení autora práce

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne Podpis autora práce

Abstrakt

Registrace oblaků bodů je důležitá úloha mobilní robotiky, která je základem simultánní lokalizace a mapování. Přínos naší práce je dvojitý: zaprvé jsme provedli podrobné porovnání lokálních registračních metod za použití vysoce kvalitních datových sad a vlastního testovacího protokolu. Naše výsledky umožňují podrobně prozkoumat vlastnosti těchto metod, zejména jejich přesnost a náchylnost na nepříznivé počáteční umístění oblaků. Zadruhé upravujeme jistou globální metodu tak, aby bylo využito vizuálních informací, konkrétně obrazu z kamer. Navrhujeme dvě vylepšení, zvýšení rozlišovací schopnosti deskriptoru a změnu algoritmu stanovení souřadných systémů klíčového bodu. Abychom ověřili kvalitu těchto úprav, vytvořili jsme datovou sadu s vizuálními daty. Výsledky našich experimentů naznačují, že došlo k výraznému zlepšení oproti původní metodě.

Abstract

Point cloud registration is an important process in mobile robotics, serving as the cornerstone of simultaneous localization and mapping. The contribution of our work is twofold: firstly, we compare local registration methods using high-quality datasets and a custom protocol. In terms of precision and robustness to initial pose displacement, the capabilities of the methods are explored in an unprecedented detail, overcoming any previous work that we know of. Secondly, we propose enhancements to a global, feature-based registration method that take advantage of visual information, specifically camera imagery. Proposed changes include an extension of the feature descriptor, and a modification of reference frame determination. To investigate the modified methods, a dataset containing visual data is created. Experimental results indicate a significant improvement over the original method.

Acknowledgements

I would like to thank my supervisor Tomáš Svoboda for supporting me and my work, and Tomáš Petříček for providing advice and insight into the subject matter. Special thanks goes to my family, and to the members of the music band Trilobajt.

Contents

1	Introduction	2
1.1	Point clouds	2
1.2	Point cloud registration	2
1.3	Local and global methods	2
1.4	Relation to the NIFTi and TRADR projects	3
2	Local methods	4
2.1	Related work	4
2.1.1	ICP: Iterative closest point	4
2.1.2	3D-NDT: Three-dimensional normal distributions transform	4
2.1.3	D2D-3D-NDT: Distribution-to-distribution three-dimensional normal distributions transform	6
2.2	Datasets	7
2.3	ETHZ protocol	8
2.4	Our protocol	9
2.5	Method composition	10
2.6	Overview of tested methods	10
2.7	Experimental results	11
2.7.1	ETHZ protocol	11
2.7.2	Our protocol	12
2.8	Conclusion	20
3	Global methods	21
3.1	Related work	21
3.2	Overview of a feature-based method	22
3.2.1	Pre-processing	22
3.2.2	Keypoint detection	22
3.2.3	Reference frame determination and disambiguation	23
3.2.4	Descriptor extraction	23
3.2.5	Descriptor matching and transformation estimation	24
3.3	Using camera imagery in feature-based registration	25
3.3.1	Camera projection and 3D gradient direction	25
3.3.2	Descriptor extraction	26
3.3.3	Reference frame determination	27
3.4	Dataset	29
3.5	Experimental results	29
3.6	Conclusion	33
4	Conclusions	34
	Bibliography	35

1 Introduction

1.1 Point clouds

Point cloud is a set of points in some coordinate system. In this work, we concentrate on three-dimensional (3D) point clouds – sets of points in 3-dimensional Euclidean space. Such point clouds can be received as the result of 3D scanning process. Common usages of 3D point clouds include surface reconstruction and object visualization. In our work, we focus on point cloud registration.

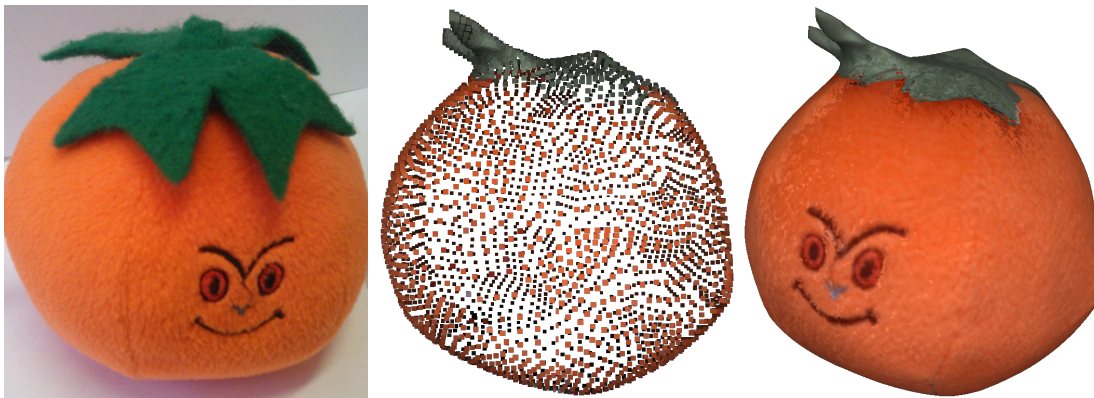


Figure 1.1 Left: photograph of an object. Middle: point cloud representation of the object received by 3D scanning. Right: 3D reconstruction of the object based on the point cloud.

1.2 Point cloud registration

Point cloud registration (a.k.a. scan registration) is the process of transforming two or more point clouds into one coordinate system, such that the corresponding overlapping parts are correctly aligned. Being recognized as an important task in multiple fields of study, such as robotics and medical imaging[7], point cloud registration has attracted widespread attention. In robotics, registration is the backbone of simultaneous localization and mapping – estimation of the robot location while simultaneously creating a global map of the surrounding environment.

Registration methods work with two point clouds; one is considered to be the *reference* or *fixed* cloud, the second is often denoted as *reading* or *moving* cloud. The distinction between the clouds is that by the registration process, the reading cloud is transformed to the coordinate system of the reference cloud. The result of a registration method is a single homogeneous transformation that converts reading data to reference coordinates.

1.3 Local and global methods

Local registration methods are based on the assumption that corresponding locations in the clouds are close, i.e. only a small transformation is needed to register the clouds.

When the relative rotation or translation of the clouds is too large, local methods will likely fail to provide the correct result. In some applications, an initial guess of the transformation is available, enabling local methods to work with large cloud displacements. This is frequently used in robotics – a transformation approximation is provided by inertial navigation systems and wheel odometry. Concerning local methods, the contribution of our work is as follows: we test a number of methods head-to-head on various difficult datasets. The results provide a clear comparison of the methods’ precision and robustness to displacement (see chapter 2).

Global registration methods function independently of the original displacement of the clouds. Therefore, in contrast to local methods, large relative translation or rotation has no effect on the result. Conventional global methods use geometrical information to find corresponding locations in the clouds. Our work contributes to this field by proposing changes to global methods that take advantage of camera imagery, i.e. visual information (see chapter 3).

1.4 Relation to the NIFTi and TRADR projects

Our work is a part of the projects NIFTi: Natural human-robot cooperation in dynamic environments, and TRADR: Long-term human-robot teaming for robot-assisted disaster response. The cornerstone of the NIFTi project is a mobile robot, designed to aid in urban search and rescue missions. The NIFTi robot features remote control, a rubber track chassis with two pairs of flippers, a rotating laser scanner, an inertial measurement unit, and an omni-directional camera.

For simultaneous localization and mapping, the robot employs odometry and a local point cloud registration method. At the time of writing, the registration method being used is iterative closest point (ICP). Our work compares the exact configuration of ICP used by the robot to legacy ICP variants, and to other local methods (see section 2.3).

Ladybug 3 omni-directional camera, mounted on top of the robot, provides visual information (e.g., color) for a majority of points provided by the laser scanner. We use the recorded point clouds and camera imagery to create a dataset for testing color-aware global methods (see section 3.4).



Figure 1.2 The NIFTi robot, featuring Sick LMS-151 laser scanner (1), and PointGrey Lady-Bug 3 omnidirectional camera (2). Images taken from [1, 4], respectively.

2 Local methods

2.1 Related work

In this section we introduce the local methods that are subject to our experiments.

2.1.1 ICP: Iterative closest point

The iterative closest point (ICP) algorithm has been originally proposed by Chen and Medioni[11], and Besl and McKay[9]. Its simplicity and ease of implementation attracted significant attention. Since its inception in 1991, a large body of related work has been created, counting over 400 papers in the past 20 years[21].

To briefly describe the algorithm: ICP iteratively improves the relative pose of two overlapping point clouds. In each iteration, the following steps are performed:

1. For each point in the reading cloud, the closest point in the reference cloud is found.
2. A transformation of the reading cloud is determined, minimizing an objective function. This is either the sum of squared distances between the corresponding points (point-to-point, as described by Besl and McKay[9]), or between a point from the reading cloud and the tangent plane of the corresponding point from the reference cloud (point-to-plane, as in Chen and Medioni[11]).
3. Found transformation is applied to the reading cloud. Tests for convergence (and divergence) of the transformation are queried, possibly ending the loop.

Finding the closest point in the reference cloud for each of the points in the reading cloud has been identified as a performance bottleneck of the method, requiring acceleration by a fast search algorithm, such as kd-tree[12]. Additionally, various pre-processing steps are applied to the clouds, for example to remove redundant data, or to pre-compute surface normals for points in the reference cloud.

Due to the popularity of ICP, a great number of its variants have emerged. To ease the selection of a proper variant for a given task, Pomerleau et al. have created the `libpointmatcher` framework[21], enabling to create and compare customized ICP configurations. Kubelka et al.[15] have proposed one such configuration, which we use in our following experiments.

The advantage of the configuration by Kubelka et. al. is that it was optimized to process real-world data scanned by the NIFTi robot. Therefore, it features a carefully constructed pipeline of point cloud pre-processing filters (see Figure 2.1), as well as a refined ICP loop (see Figure 2.2). In section 2.3 we introduce a series of tests involving this version of ICP, as well as the legacy point-to-point and point-to-plane variants.

2.1.2 3D-NDT: Three-dimensional normal distributions transform

The three-dimensional normal distribution transform (3D-NDT) algorithm has been described thoroughly by Magnusson[17], extending the original normal distribution transform by Biber and Straßer[10]. At the core of these methods is the intention to create a different representation for point cloud data. Using normal distributions transform, groups of points are used to calculate normal distributions, creating a statistical model. To extract the parameters μ , Σ of a normal distribution for a group of

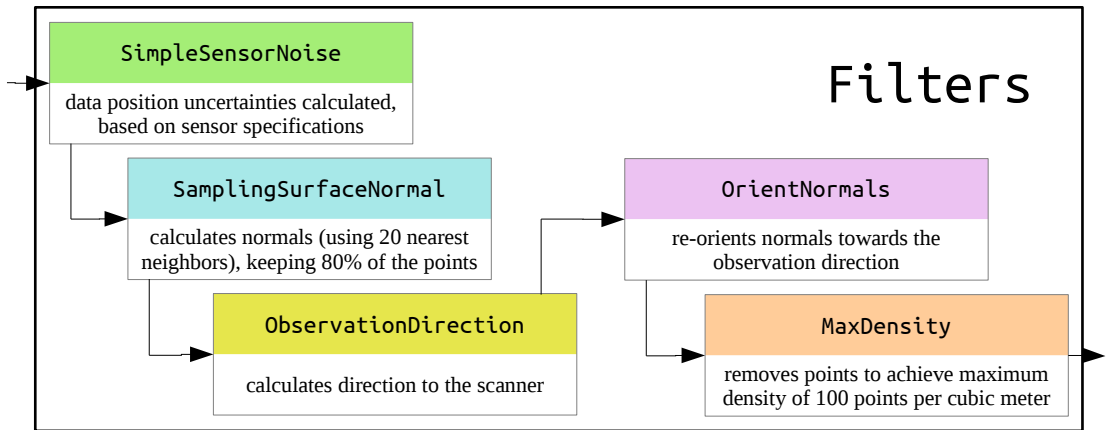


Figure 2.1 Diagram of pre-processing filters used on input clouds by Kubelka et al.[15]

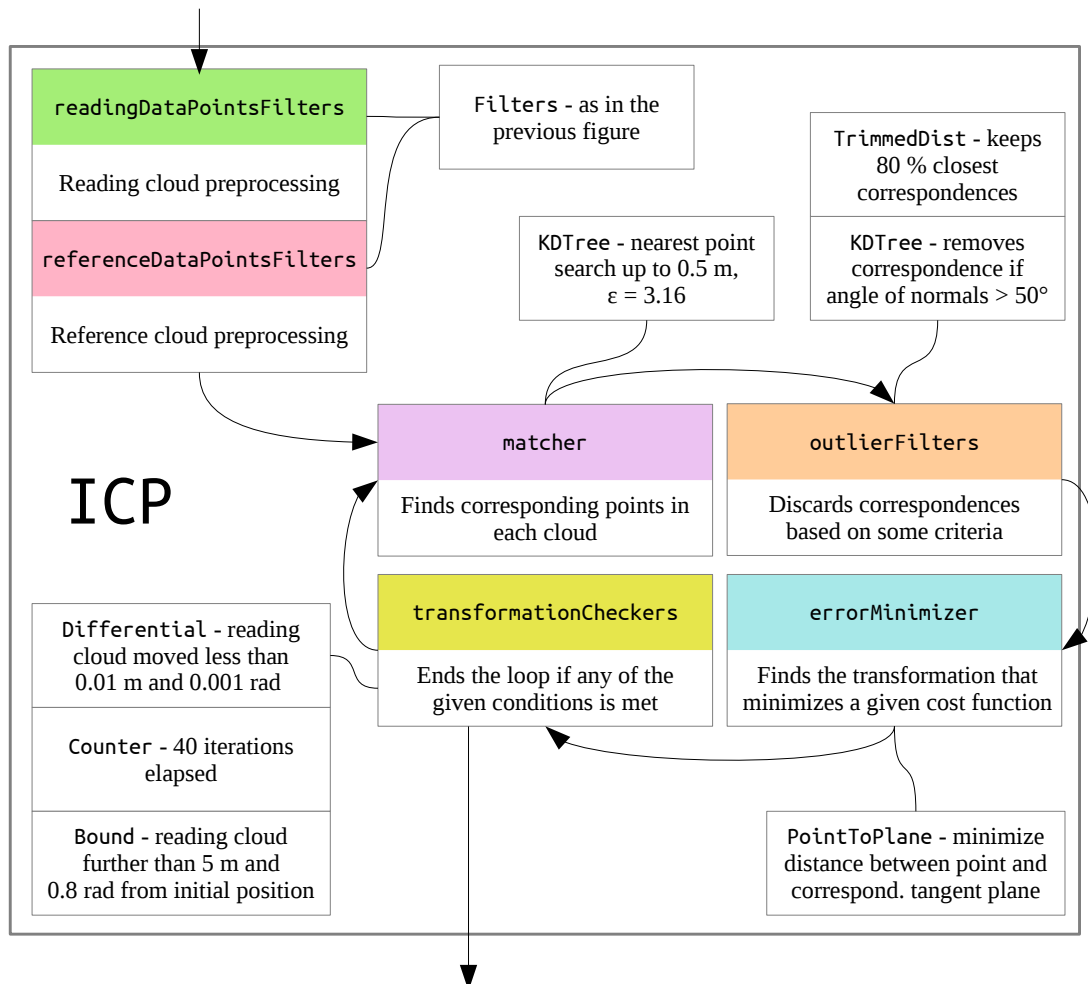


Figure 2.2 Diagram of the ICP algorithm key steps, as implemented in the libpointmatcher library[21] and configured by Kubelka et al.[15]

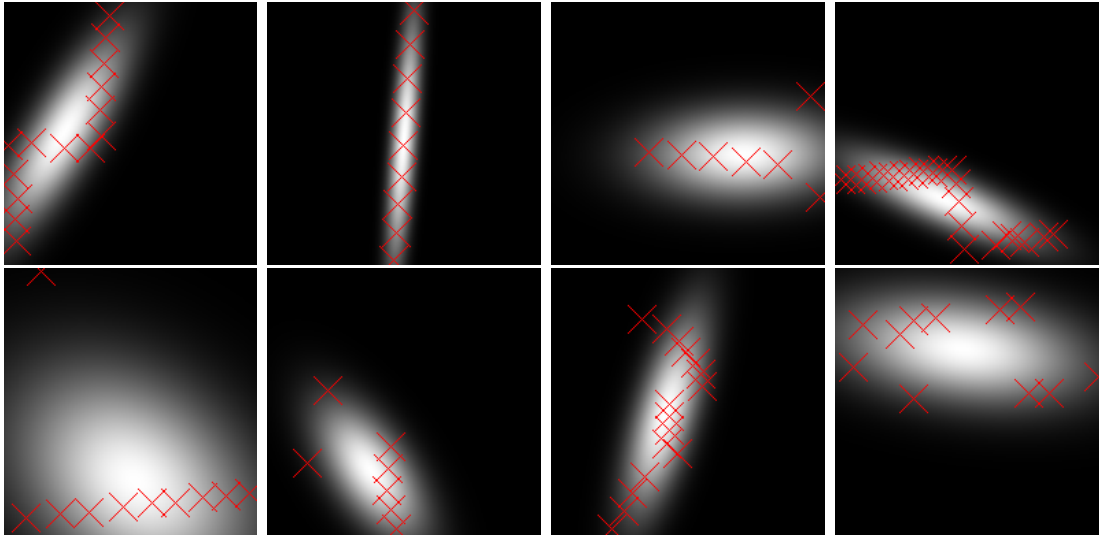


Figure 2.3 Normal distributions transform. In a plane, groups of points (red crosses) are converted into normal distributions (grey background). Drawing taken from [10].

points $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n$, maximum likelihood estimation is employed:

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{p}_i \quad (2.1)$$

$$\boldsymbol{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{p}_i - \boldsymbol{\mu})(\mathbf{p}_i - \boldsymbol{\mu})^T \quad (2.2)$$

The resulting representation is a compact model of the scanned surfaces, with applications beyond the scope of point cloud registration. Consequently, the term normal distributions transform refers to the process of converting points into normal distributions, as well as to the registration method that makes use of this process.

Similarly to ICP, the 3D-NDT registration method refines the relative pose of the clouds iteratively, and by maximizing an objective function. During initialization, points in the reference cloud are replaced by Gaussians, while the reading cloud is left unchanged. In an iteration step, the objective function being maximized expresses the likelihood that reading points were generated by their respective nearest of the reference cloud's distributions.

Examples of known extensions of the method include: Color-NDT[13], making use of visual information; trilinear interpolation, bringing eight nearest distributions into consideration; and D2D-3D-NDT, which we shall introduce next.

2.1.3 D2D-3D-NDT: Distribution-to-distribution three-dimensional normal distributions transform

The distribution-to-distribution three-dimensional normal distribution transform (D2D-3D-NDT), proposed by Stoyanov et al.[24], is a local method related closely to 3D-NDT. To distinguish the two, Stoyanov et al. refer to the original 3D-NDT as the point-to-distribution variant.

In contrast to the point-to-distribution method, D2D-3D-NDT converts both clouds, reading and reference, into Gaussians. The minimized objective function is based on L_2 distances between pairs of closest distributions from each cloud.

In the experiments presented by Stoyanov et al., the distribution-to-distribution approach has shown promising results. Our work thoroughly compares this method to 3D-NDT and others, providing an overview of its precision and robustness to initial displacement.

2.2 Datasets

Datasets used in our local method experiments were introduced by Pomerleau et al. [22]. These eight point cloud sequences cover a diverse range of challenging environments and

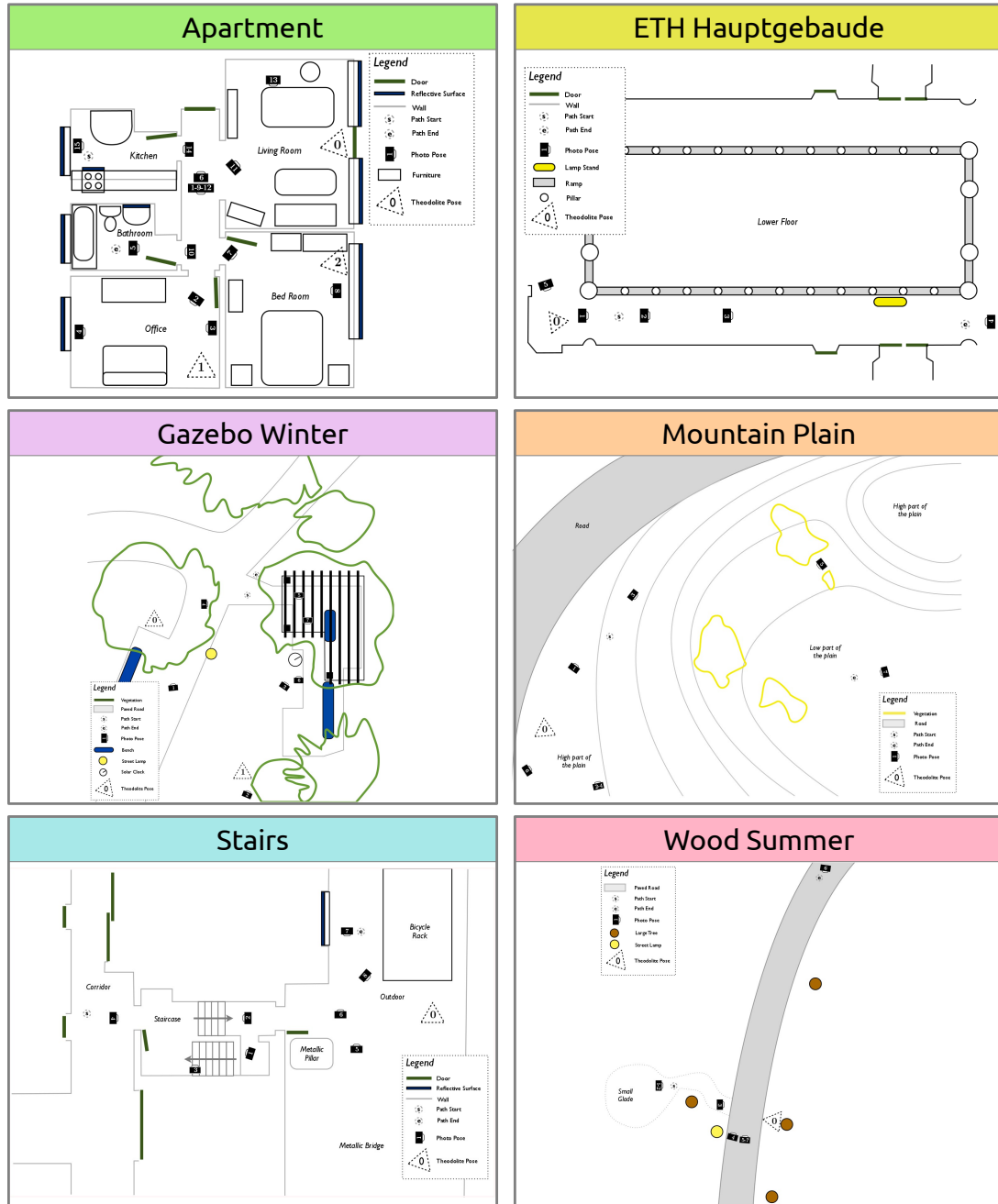


Figure 2.4 Top-down maps of the datasets. Drawings taken from [22].

situations, providing precise ground truth poses of the sensor measured by a theodolite (translation error under 1.8 mm, rotation error under 0.006 rad[22]). Hokuyo UTM-30LX laser scanner was used to capture the data, providing 100.000 to 350.000 points per scan. Let us describe the six sequences that we used in our experiments:

- Apartment – a sequence of point clouds in a single-floor apartment with five rooms: a kitchen, a bathroom, an office and a bedroom. Such environment is well structured, i.e. most surfaces in the environment are representable by geometric primitives. Although some dynamic elements were created by purposely relocating objects between scans, the apartment sequence is considered the least difficult to register.
- ETH Hauptgebäude – point clouds in this sequence were scanned in a hallway, featuring repetitive elements, such as pillars and arches. This creates an interesting challenge, as local registration methods tend to converge to the local minima of their respective cost function, which repetitive elements may create.
- Gazebo Winter – an outdoor, semi-structured dataset, with both geometrically simple and complex surfaces. The point clouds were scanned near a summer house in a park, surrounded by trees. To further increase registration difficulty, people were recorded both sitting and in motion – walking in the scene during the scanning process.
- Mountain Plain – an unstructured outdoor scene, featuring no man-made structures and no obvious vertical landmarks. Covered in approximately 0.5 m tall grass, the plain has proven to be a substantially challenging scene for registration methods, as there are no vertical surfaces to sufficiently constrain the registration process.
- Stairs – a dataset for testing methods’ robustness to large changes in scanned volumes, i.e. sizes of areas represented by a point cloud. In the sequence, starting indoors, the scanner first passes through a few doorways, eventually leaving the building into an outdoor environment.
- Wood Summer – an outdoor scene, recorded in a forest. Apart from a small paved road, all objects in the scene (trees and other vegetation) consist of unstructured surfaces. Furthermore, as in the Gazebo Winter dataset, dynamic elements were created by recording people in motion.

Hand in hand with a dataset is a *protocol* – a pre-generated sequence of scan pairs to register, along with initial guesses of the relative pose. In our experiments, we use two different protocols for the above-mentioned datasets. On these protocols, we elaborate in detail in sections 2.3 and 2.4.

2.3 ETHZ protocol

First of the two protocols we used to test local methods is one proposed by Pomerleau et al.[21], from ETH Zürich, based on the datasets described in the previous section. The protocol features 35 pairs of clouds for each of the six datasets, with 192 initial poses for each pair, totalling over 40,000 registrations. For a dataset, point cloud pairs have been selected so that their overlaps, i.e. the amount of represented surfaces that are common to both clouds, are distributed approximately uniformly from 30 % to 99 %.

Each protocol entry consists of a reading and a reference cloud, an initial transformation for the reading cloud, and an expected resulting transformation T_G (the ground truth). To evaluate the protocol, the method in question is run on all the entries. After a method registers a pair of clouds, the resulting transformation T_R is analyzed to

receive translational error ϵ_t and rotational error ϵ_r :

$$\Delta T = T_R T_G^{-1} = \begin{bmatrix} \Delta R & \Delta t \\ 0 & 1 \end{bmatrix} \quad (2.3)$$

$$\epsilon_t = \|\Delta t\|_2 \quad (2.4)$$

$$\epsilon_r = \arccos \left(\frac{\text{trace}(\Delta R) - 1}{2} \right) \quad (2.5)$$

The above formula extracts the minimum distance (in meters) and minimum angle (in radians), by which the final pose T_R is be moved to be identical to the ground truth pose T_G . To compare method capabilities, we use three quantiles of rotational and translational errors:

A50	0.5-quantile, the median
A75	0.75-quantile
A95	0.95-quantile

Initial poses were artificially generated from the ground truth pose by offsetting it by a perturbation (i.e. displacement). Three perturbation types are used in this protocol, Easy, Medium and Hard, with increasing standard deviations of zero-mean Gaussians, from which samples were taken to generate the displacements. For each perturbation type and point cloud pair, the protocol contains 64 initial poses.

We evaluate this protocol to explore the precision and robustness of three new methods (ICP by Kubelka et al.[15], 3D-NDT, D2D-3D-NDT), in comparison to one another and to legacy ICP variants (see section 2.6 for an overview of the tested methods). Results for the Easy perturbation type represent the situation where the method was given a good initial pose; the lower the error for A50 and A75 quantiles, the better the precision of the method. On the other hand, low A95 quantiles are generally an indication of method robustness. Furthermore, results may vary greatly for different datasets, given their diverse nature (see section 2.2).

2.4 Our protocol

Although the ETHZ protocol is sufficient for basic method comparison, we chose to create our own protocol in order to explore the *limitations* of local registration methods. We suspect that methods are variously susceptible to the two types of initial displacement, translational and rotational, and their combinations. Therefore, we would like to investigate a larger number of perturbation types than the three of the ETHZ protocol.

Our protocol resembles the ETHZ protocol in most of the features. It is based on the same datasets (see section 2.2), and is using the same point cloud pairs. On the other hand, we increased the number of perturbation types from 3 to 25, these being combinations of 5 translational and 5 rotational perturbation types. The following table lists standard deviations of zero-mean Gaussians, from which samples were taken to generate displacement:

α	[rad]	d	[m]
R1	0.0625	T1	0.125
R2	0.1250	T2	0.250
R3	0.2500	T3	0.500
R4	0.5000	T4	1.000
R5	1.0000	T5	2.000

To generate a protocol entry, given its perturbation type, an angle sample α and a distance sample d were taken from the corresponding distributions. To displace the ground truth pose rotationally, we rotate it by α about a random axis; to perform the translational displacement, we translate it by d in a random direction. The final, perturbed pose is used as the initial pose for registration. As in the ETHZ dataset, 64 initial poses were generated for each point cloud pair and perturbation type.

By evaluating this protocol, we push the tested methods to their limits, exploring their robustness to a large number of combinations of rotational and translational displacements. This allows us to investigate the methods in unprecedented detail, overcoming any previous work on the matter. For all datasets, we find the limitations of the tested algorithms, i.e. our results indicate a maximum displacement for a method to operate, given error requirements. Furthermore, by providing a detailed comparison of the methods, the experimental results are a valuable resource for finding the best algorithm for a given environment and task.

2.5 Method composition

Intuitively, a method that is generally robust to initial displacement can be composed with a precise method to form a new, composite method that is potentially superior to its parts. Our initial experiments with the ETHZ protocol (these experimental results are demonstrated in section 2.7.1) suggested that while D2D-3D-NDT falls short to ICP and 3D-NDT in terms of robustness, its capabilities of precise registration were more than satisfactory. Therefore, we decided to compose D2D-3D-NDT as the back end to ICP and 3D-NDT, creating two chained methods. In order to investigate the feasibility of these methods, we tested them using both experimental protocols. The results are shown in sections 2.7.1 and 2.7.2.

2.6 Overview of tested methods

Here we provide a summary of the tested methods.

- Besl ICP, Chen ICP – legacy point-to-point and point-to-plane ICP methods, as described by Besl and McKay[9] and Chen and Medioni[11]. Evaluated results of these methods were provided by Pomerleau et. al.[21]. ICP is discussed in detail in section 2.1.1. In our experiments, these legacy methods establish a baseline to which one can compare the following new algorithms.
- ICP – an ICP configuration by Kubelka et al.[15]. We describe the algorithm in diagram 2.2. As an example of a configuration refined on a real-world registration application, we anticipate its capabilities to be of the best currently attainable by a variant of ICP.
- P2D – an implementation of 3D-NDT from the PCL library, publicly available at [5]. The reading cloud is filtered as in the ICP method (see diagram 2.1). For more information on 3D-NDT, see section 2.1.2.
- D2D – an implementation of the D2D-3D-NDT method provided by Center for Applied Autonomous Sensor Systems at Örebro University, Sweden, publicly available at <http://code.google.com/p/oru-ros-pkg/>. We describe this method in section 2.1.3.
- ICP-D2D, P2D-D2D – composite methods. We explain our motivation to include these methods in section 2.5.

2.7 Experimental results

2.7.1 ETHZ protocol

Below are the results obtained by evaluating the protocol described in section 2.3. In the following tables, rows correspond to tested methods (as in section 2.6), and columns correspond to datasets (as in 2.2). For each dataset and method, three quantiles (A50, A75, A95) of two error types are shown; the upper part of a table contains rotation error in radians from (2.5), the lower part contains translation error in meters. Results are color-coded by dataset, with more erroneous results highlighted with a more saturated shade of a color. One table is presented for each perturbation type, Easy, Medium and Hard. The best result for a given perturbation type, dataset and quantile is in bold. These are our observations based on the results below:

- ICP and P2D provide similar results. These methods seem to be the best in terms of robustness (compare ICP and P2D to other methods, all perturbation types, all datasets, A95). In comparison, P2D seems slightly more precise and robust.
- D2D and ICP-D2D provide similar results as well. Although evidently not very robust (compare D2D to ICP, Medium perturbation type, all datasets, A95), their precision appears to be quite good (compare D2D to ICP, Medium perturbation type, all datasets, A50), especially on Gazebo Winter and Wood Summer datasets.
- P2D-D2D has performed poorly, with all results inferior or comparable to P2D. Although its A95 quantiles are satisfactory in Easy and Medium perturbation types, its precision is severely lacking (see A50 quantiles of P2D-D2D in Easy and Medium perturbation types).

Easy perturbation type

dataset		Apartment			ETH			Gazebo			Plain			Stairs			Wood		
quantile		A50	A75	A95	A50	A75	A95	A50	A75	A95	A50	A75	A95	A50	A75	A95	A50	A75	A95
Rotation	ICP Besl	0.07	0.25	0.97	0.05	0.22	0.83	0.04	0.17	0.41	0.09	0.20	0.44	0.12	0.39	1.22	0.09	0.29	0.77
	ICP Chen	0.02	0.20	1.14	0.01	0.02	0.61	0.02	0.08	0.48	0.07	0.20	0.60	0.02	0.31	1.58	0.05	0.34	0.95
	ICP	0.01	0.02	0.21	0.00	0.01	0.34	0.02	0.03	0.32	0.01	0.02	0.18	0.01	0.01	0.33	0.01	0.20	0.43
	P2D	0.02	0.02	0.24	0.01	0.13	0.42	0.02	0.02	0.30	0.02	0.17	0.38	0.01	0.03	0.37	0.01	0.04	0.40
	D2D	0.02	0.11	0.65	0.00	0.01	0.43	0.01	0.02	0.15	0.09	0.32	3.03	0.01	0.16	1.06	0.01	0.02	0.37
	ICP-D2D	0.02	0.07	0.63	0.00	0.01	0.45	0.01	0.02	0.15	0.08	0.31	3.08	0.01	0.11	0.58	0.01	0.02	0.36
	P2D-D2D	0.27	0.39	0.49	0.21	0.30	0.43	0.23	0.32	0.44	0.23	0.33	0.47	0.24	0.34	0.47	0.22	0.32	0.43
Translation	ICP Besl	0.13	0.54	1.54	0.47	2.23	6.86	0.28	0.60	1.71	0.51	1.46	3.09	0.35	1.29	2.57	0.39	1.48	4.21
	ICP Chen	0.06	0.47	2.11	0.10	0.44	6.06	0.11	0.38	2.08	0.42	1.54	4.15	0.09	1.17	3.49	0.25	1.55	4.75
	ICP	0.03	0.04	0.36	0.02	0.03	2.11	0.03	0.08	0.69	0.06	0.15	1.16	0.03	0.05	0.94	0.04	0.35	1.01
	P2D	0.03	0.04	0.37	0.19	0.47	1.22	0.03	0.07	0.66	0.11	0.38	0.89	0.03	0.08	0.73	0.04	0.14	0.91
	D2D	0.03	0.12	1.06	0.04	0.20	2.22	0.04	0.08	0.25	0.14	0.27	4.16	0.03	0.17	1.84	0.07	0.15	0.49
	ICP-D2D	0.04	0.12	1.08	0.04	0.22	2.22	0.04	0.10	0.25	0.16	0.33	4.03	0.04	0.15	1.83	0.08	0.18	1.48
	P2D-D2D	0.17	0.23	0.65	0.21	0.37	1.22	0.19	0.27	0.64	0.22	0.35	3.11	0.18	0.25	0.84	0.19	0.28	0.78

Medium perturbation type

dataset		Apartment			ETH			Gazebo			Plain			Stairs			Wood		
quantile		A50	A75	A95	A50	A75	A95	A50	A75	A95	A50	A75	A95	A50	A75	A95	A50	A75	A95
Rotation	ICP Besl	0.20	0.61	1.49	0.14	0.59	1.82	0.15	0.35	0.80	0.20	0.37	0.77	0.33	0.78	1.63	0.32	0.69	1.22
	ICP Chen	0.08	0.47	1.80	0.01	0.25	2.91	0.04	0.35	0.97	0.19	0.38	0.99	0.16	1.08	2.09	0.31	0.78	1.53
	ICP	0.07	0.71	1.14	0.01	0.64	1.10	0.34	0.64	1.04	0.06	0.68	1.00	0.07	0.71	1.20	0.48	0.75	1.05
	P2D	0.23	0.64	1.12	0.45	0.70	1.10	0.33	0.62	1.02	0.25	0.48	0.98	0.29	0.69	1.16	0.47	0.72	1.05
	D2D	0.02	0.55	2.04	0.01	0.54	1.62	0.01	0.11	1.20	0.18	0.74	3.09	0.01	0.70	1.98	0.01	0.48	1.13
	ICP-D2D	0.03	0.84	1.95	0.01	0.68	1.40	0.01	0.04	1.20	0.38	0.86	3.03	0.02	0.81	1.70	0.01	0.47	1.09
	P2D-D2D	0.43	0.66	1.38	0.30	0.48	1.15	0.31	0.45	0.93	0.48	0.74	1.14	0.40	0.62	1.55	0.27	0.49	1.02
Translation	ICP Besl	0.46	1.03	2.32	1.92	4.29	11.24	0.49	1.13	3.18	1.21	2.17	3.76	0.94	1.86	3.38	1.19	2.52	5.15
	ICP Chen	0.20	1.04	2.98	0.60	4.06	16.26	0.28	0.96	3.51	1.30	2.58	5.58	0.61	2.08	4.64	1.25	2.92	6.62
	ICP	0.31	0.95	1.59	0.71	1.80	3.03	0.62	1.10	1.79	0.35	0.98	2.11	0.60	1.17	1.98	0.79	1.22	1.88
	P2D	0.54	0.99	1.55	0.86	1.33	2.03	0.60	1.00	1.57	0.64	1.04	1.63	0.63	1.10	1.71	0.68	1.12	1.75
	D2D	0.04	0.94	2.23	0.51	1.61	3.39	0.05	0.30	1.59	0.53	1.18	4.08	0.24	1.31	3.14	0.15	0.98	3.88
	ICP-D2D	0.18	0.96	2.04	0.51	1.55	4.19	0.06	0.27	1.73	0.64	1.10	3.10	0.26	1.09	2.76	0.17	1.07	3.59
	P2D-D2D	0.70	1.02	1.95	0.97	1.44	3.00	0.74	1.00	1.55	0.74	1.13	3.53	0.81	1.18	2.78	0.80	1.16	3.23

Hard perturbation type

dataset quantile		Apartment			ETH			Gazebo			Plain			Stairs			Wood		
		A50	A75	A95	A50	A75	A95	A50	A75	A95	A50	A75	A95	A50	A75	A95	A50	A75	A95
Rotation	ICP Besl	1.04	1.60	2.53	0.97	1.73	3.05	0.58	1.20	2.59	0.46	0.99	2.09	1.10	1.64	2.53	0.97	1.44	2.35
	ICP Chen	1.01	1.72	2.95	1.31	2.09	3.11	0.58	1.31	2.88	0.50	1.09	3.05	1.48	1.91	2.94	1.05	1.56	2.53
	ICP	1.15	1.58	2.15	1.10	1.53	2.13	1.05	1.46	2.09	1.13	1.42	2.08	1.14	1.57	2.11	1.08	1.46	2.10
	P2D	1.10	1.52	2.12	1.07	1.46	2.04	1.03	1.43	2.10	0.80	1.36	2.27	1.09	1.50	2.10	1.08	1.45	2.08
	D2D	1.04	1.78	3.11	0.95	1.60	3.01	0.97	1.54	2.90	1.05	1.75	3.12	1.15	1.79	3.05	1.02	1.50	2.26
	ICP-D2D	1.14	1.64	2.99	1.02	1.57	2.59	0.97	1.54	2.82	1.15	1.52	2.38	1.13	1.61	2.85	1.02	1.49	2.38
	P2D-D2D	0.94	1.65	2.83	0.95	1.50	2.57	0.92	1.44	2.47	1.01	1.52	2.82	0.96	1.63	2.82	1.02	1.45	2.15
Translation	ICP Besl	1.29	1.99	3.24	3.84	7.06	14.77	1.58	2.79	4.57	2.02	3.14	6.33	1.81	2.78	4.75	2.32	3.73	6.82
	ICP Chen	1.35	2.18	3.66	4.18	8.55	19.56	1.87	3.33	6.95	2.35	4.13	8.85	2.05	3.28	5.50	2.79	4.52	7.86
	ICP	1.49	2.05	2.80	1.85	2.68	4.25	1.64	2.27	3.21	1.47	2.06	2.72	1.63	2.32	3.08	1.64	2.25	3.07
	P2D	1.48	2.03	2.76	1.63	2.22	3.04	1.52	2.04	2.85	1.43	1.98	2.72	1.58	2.14	2.86	1.54	2.08	2.86
	D2D	1.13	2.00	3.04	1.83	2.68	5.16	1.37	2.32	3.90	1.42	2.32	4.31	1.58	2.58	4.08	1.47	2.57	5.41
	ICP-D2D	1.37	2.07	2.85	1.71	2.58	4.99	1.32	2.31	3.95	1.47	2.09	2.78	1.47	2.37	3.69	1.51	2.64	5.44
	P2D-D2D	1.23	1.93	3.09	1.74	2.62	4.83	1.34	2.00	3.58	1.48	2.21	4.07	1.51	2.46	3.96	1.50	2.48	4.71

2.7.2 Our protocol

Below are the results obtained by evaluating the protocol described in section 2.3. On the following pages, there are two tables for each dataset, showing rotation error in radians from (2.5) and translation error in meters. In a table, rows correspond to tested methods (see section 2.2), while columns correspond to error quantiles (A50, A75, A95). For each method and quantile, results are presented in a small table, where columns correspond to rotational perturbation types (R1, . . . , R5), and rows correspond to translational perturbation types (T1, . . . , T5). As before, results are color-coded by dataset, with more erroneous results highlighted with a more saturated shade of a color. The best result for a given perturbation type, dataset and quantile is in bold. These are our observations based on the results below:

- P2D provides results largely similar to ICP. In terms of precision, ICP performs slightly better overall, especially on the ETH Hauptgebäude dataset (compare ICP to P2D, A50). On the other hand, P2D shows an increase of robustness in some situations, particularly in the case of large rotational displacement (compare ICP to P2D, translational error, A95, R5, on ETH Hauptgebäude, Mountain Plain and Stairs datasets).
- D2D and ICP-D2D provide nearly identical results; both methods seem to have the same strengths and shortcomings. On severely unstructured, foliage-laden datasets (Gazebo Winter, Mountain Plain, Wood Summer) we see some error decrease relative to ICP and P2D, but only for specific perturbation types (compare D2D to ICP, translational error, A75, R4). Overall, precision when given a good initial position suffers greatly (compare ICP to D2D, A50, R1), as well as robustness (compare ICP to D2D, A95, R1). The superior precision of D2D suggested by the ETHZ protocol results (see section 2.7.1) is not apparent here.
- P2D-D2D, while it does improve on the robustness of D2D (compare D2D to P2D-D2D, rotational error, A95, on Mountain Plain dataset), comes out as the worst method in terms of precision on any dataset (compare P2D-D2D to any other method, A50, R1).

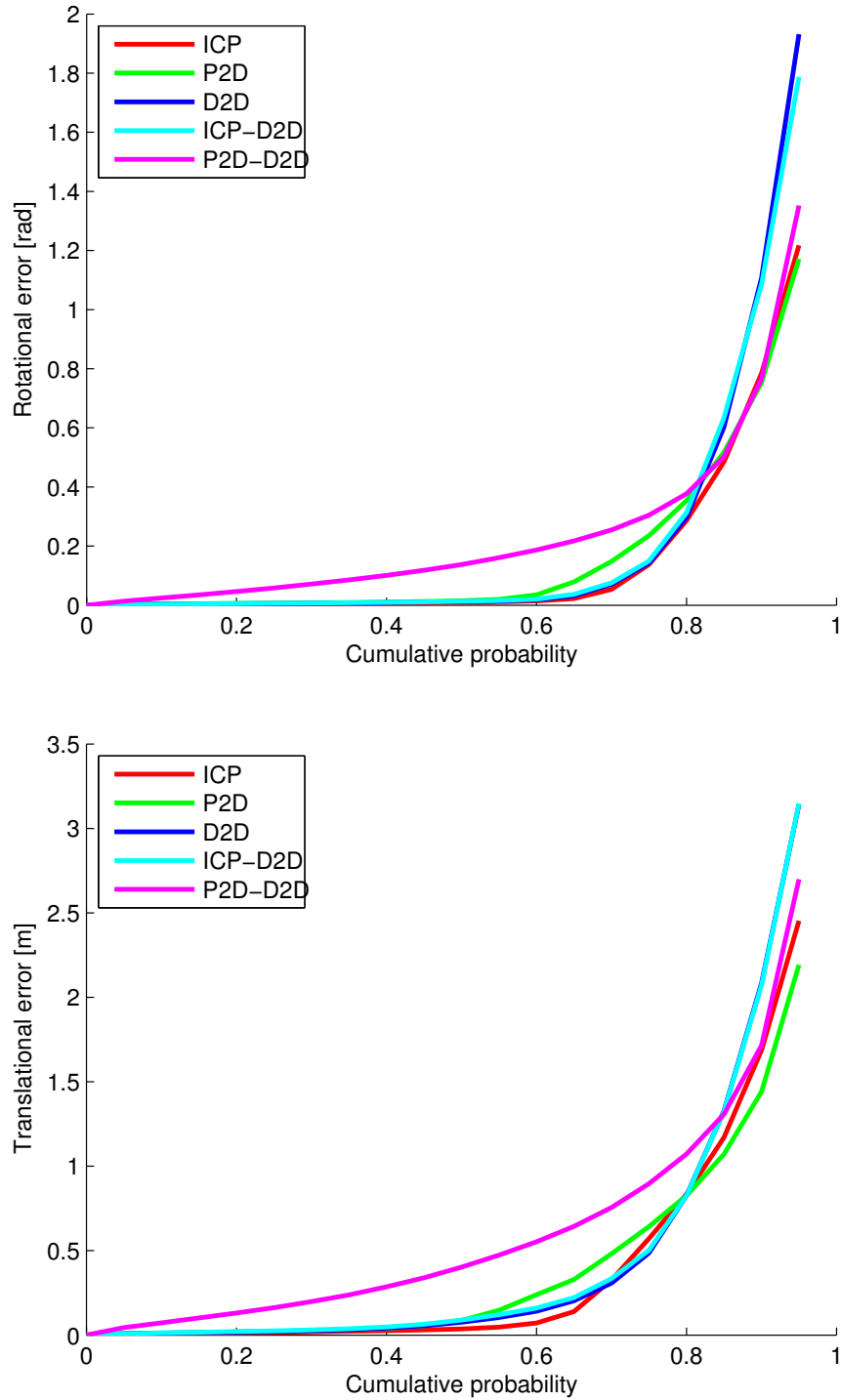


Figure 2.5 A condensation of results generated by our protocol: quantiles of rotational and translational error over all datasets and perturbation types. P2D-D2D is shown lacking in low quantiles, suggesting low precision. D2D and ICP-D2D show nearly identical results, lacking in high quantiles, signifying decreased robustness. ICP and P2D perform comparably, with ICP having an edge in precision, while P2D excels in robustness.

2.8 Conclusion

Using a custom, detailed protocol based on well-established datasets, we provide an unprecedented level of insight into precision and robustness of local methods, overcoming any previous work that we know of.

Analysis of our data provides information that simplifies selection of the best method for a given environment. Specifically, our results suggest an overall dominance of the ICP method configured by Kubelka et al.[15], and the Point Cloud Library implementation of the 3D-NDT (P2D) method. On any of the diverse set of datasets, one of these two methods provides the best results, both robustness- and precision-wise. Therefore, we suggest these methods to be the first choice for a 3D point cloud registration task.

Furthermore, using our results, one can find the limitations of any of the tested methods. Given error requirements, one can identify a maximum displacement that the method can overcome in a specific environment. This gives a clear picture of a method's capabilities with chosen constraints on translational and rotational error. Additionally, when evaluated using our protocol, any other algorithms and configurations can be thoroughly compared to the methods included in our experiments.

Needless to say, susceptibility to displacement of the initial pose is a fundamental limitation of all local methods. To perform registration with an arbitrary initial pose, a global method is needed.

3 Global methods

In contrast to local methods, global registration methods function independently of the initial relative pose of the clouds. Consequently, no initial guess of the relative pose is needed, as cloud displacement has no effect on the result. In our work, we focus on *feature-based* global registration methods. By *features* we mean descriptions of local point cloud data, which can be extracted and matched with each other. Features are created at *keypoints*, i.e. points of interest in a cloud. The core problem of feature-based registration is the maximization of *repeatability* – the same keypoints and features are to be found in different point clouds. Repeatability enables us to find corresponding features in the reading and reference clouds, which we use to estimate the relative pose.

3.1 Related work

The Spin images method, proposed by Johnson and Herbert[14], is a widely used approach to feature-based registration. This method establishes a reference axis at each keypoint – typically, in the direction of an estimated surface normal. About this axis, a plane is rotated, and intersections of nearby points with the plane are marked into a “spin image” (see figure 3.1) – a 2D histogram which is used as a feature descriptor for matching.

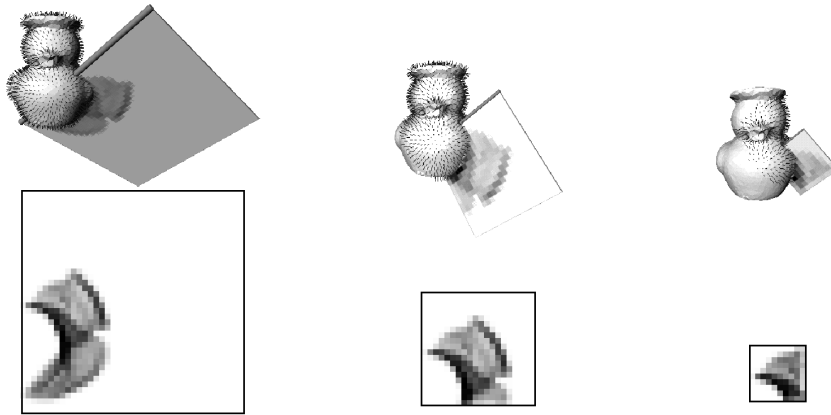


Figure 3.1 Spin images of different sizes are created for a keypoint. A plane rotates about a surface normal (top), marking intersections with nearby points into 2D histograms (bottom). Drawings taken from [14].

Fast Point Feature Histogram (FPFH) is a modern descriptor by Rusu et al.[23] Instead of relying on a reference axis, FPFH analyzes rotationally-invariant geometrical relations among points near the keypoint. Another descriptor, Signature of Histograms of Orientations (SHOT), was proposed by Tombari et al.[25] In contrast to the single reference axis of Spin, SHOT is a descriptor that requires to establish a reference *frame*, i.e. three orthogonal axes. Features aligned to reference frames generally excel in discriminative power of their descriptors, although as noted in [25], the task of repeatably finding the reference frames is crucial in this case.

The feature-based method by Petricek and Svoboda[19] also employs a reference frame-based descriptor[20]. Our work is an extension of [19], proposing changes to its descriptor and reference frame determination that take advantage of visual information. A large body of work has been created on the topic of feature-based image registration, based solely on visual data. Prominent methods in this area include Scale Invariant Feature Transform (SIFT) by Lowe and David[16] and Speeded Up Robust Features (SURF) by Bay et al.[8] Binary Robust Appearance and Normals Descriptor (BRAND) by Nascimento et al.[18] is an example of a descriptor that fuses range and visual data, taking depth information from RGB-D images into consideration.

3.2 Overview of a feature-based method

In this section, we describe the key steps of a feature-based registration algorithm. Additionally, we discuss the specific implementations of these steps in the method by Petricek and Svoboda[19] that we propose modifications to in section 3.3.

3.2.1 Pre-processing

As in local methods (e.g. see figure 2.1), a sequence of filters is first applied to each of the clouds being registered to remove redundant and erroneous data, and to calculate additional properties of points. Most feature-based methods suffer from non-uniform sampling density [14, 25, 19], requiring a density filter, and many methods also make use of surface normals calculated for each point in pre-processing [23, 25, 19].

3.2.2 Keypoint detection

Determining the locations of features is accomplished by defining a saliency measure, i.e. a measure of interest in a given location. Each point in a cloud is treated as a keypoint candidate; the saliency measure is calculated at the position of every point, and points that produce local maxima of the measure are then promoted to keypoints. To calculate the measure in [19] for a given point, nearby points \mathbf{p}_i are found up to the distance of 0.35 m. Then, the covariance matrix C of point positions is found:

$$\boldsymbol{\mu} = \frac{\sum_i \mathbf{p}_i}{\sum_i 1} \quad (3.1)$$

$$C = \sum_i (\mathbf{p}_i - \boldsymbol{\mu})(\mathbf{p}_i - \boldsymbol{\mu})^T \quad (3.2)$$

As C is a positive semidefinite matrix with real coefficients, singular value decomposition (SVD) of C yields real positive eigenvalues $\lambda_1 > \lambda_2 > \lambda_3$ and corresponding orthonormal eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$:

$$C = USV^T = VSV^T = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \mathbf{v}_3] \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \mathbf{v}_3^T \end{bmatrix} \quad (3.3)$$

In [19], it is shown that λ_3 is a good choice of a saliency measure; flat surfaces are given a low saliency score in favor of edges and corners. Finally, non-maximum suppression is used to find the local maxima – a keypoint candidate is discarded if there is a more salient point in a 0.2 m radius. Remaining candidates form keypoints at the mean positions $\boldsymbol{\mu}$ from equation 3.1.

3.2.3 Reference frame determination and disambiguation

This step is performed only for methods that require a reference frame to align their descriptor [25, 20]; some descriptors require only a single reference axis[14] or no reference axes at all[23]. Tombari et al.[25] stress the importance of repeatable determination of reference frames, which they believe is underrated in favor of descriptor choice.

To determine the reference frame of a given keypoint, Petricek and Svoboda[19] take the principal components of positions of nearby points found within a radius of 2 m. A covariance matrix C is created from these points (as in equations 3.1, 3.2) and SVD applied (as in 3.3) – eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ then form principal components, with \mathbf{v}_1 in the direction of the largest position variance, and \mathbf{v}_3 in the direction of the lowest.

Directions of eigenvectors of the point covariance matrix are repeatable and widely used as the reference frame of a keypoint [25, 19]; unfortunately, their signs are determined accidentally by the SVD, creating four ambiguous rotation possibilities. Some feature-based methods alleviate the problem by extracting multiple descriptors for a keypoint, one for each of the ambiguous reference frames. Tombari et al.[25] instead propose a sign disambiguation method, improving the repeatability of the signs. The sign disambiguation method by Petricek and Svoboda[19] forces the orientation of two axes towards the position of the scanner. Given eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, a keypoint location $\boldsymbol{\mu}$, and a scanner location \mathbf{s} , the reference frame $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ is calculated as follows:

$$\tau(\mathbf{x}) = \begin{cases} 1 & \text{if } (\mathbf{s} - \boldsymbol{\mu})^T \mathbf{x} \geq 0 \\ -1 & \text{if } (\mathbf{s} - \boldsymbol{\mu})^T \mathbf{x} < 0 \end{cases} \quad (3.4)$$

$$\mathbf{a}_2 = \tau(\mathbf{v}_2) \cdot \mathbf{v}_2 \quad \mathbf{a}_3 = \tau(\mathbf{v}_3) \cdot \mathbf{v}_3 \quad \mathbf{a}_1 = \mathbf{a}_2 \times \mathbf{a}_3 \quad (3.5)$$

In [19], it is noted that \mathbf{a}_3 provides a very repeatable direction, including the sign. With \mathbf{v}_3 being the direction of the lowest variance of positions, it is an estimation of the direction of the surface normal. All visible surfaces are inherently oriented towards the scanner; consequently, the sign disambiguation method based on sensor location consistently enforces the correct orientation of \mathbf{a}_3 . On the other hand, the directions of $\mathbf{v}_1, \mathbf{v}_2$ are not as stable, susceptible to missing parts of the surface (due to occlusion) and to non-uniform sampling density. To create a right-handed coordinate system, \mathbf{a}_1 is calculated as the cross product of \mathbf{a}_2 and \mathbf{a}_3 .

3.2.4 Descriptor extraction

The feature descriptor is an essential part of any feature-based method. Descriptors are designed to strike the balance between discriminative power and descriptor size, which affects memory requirements and performance. Most methods employ a histogram as their descriptor [14, 23, 25, 19], while some rely on a binary string [18].

In [19], Petricek and Svoboda use a histogram of point positions and normal directions as proposed in [20]. To extract the descriptor, nearby points up to the distance of 2 m are considered, along with their corresponding pre-calculated normals. Let us describe the process of marking a given point into the histogram (see figure 3.2).

- (a) A point with a corresponding normal is located near the keypoint.
- (b) Both the point and its normal are orthogonally projected onto one of the three planes determined by the axes of the reference frame.
- (c) The position of the point is weighted into a four-bin spatial histogram using linear interpolation.

(d) The direction of the normal is weighted into an eight-bin angular histogram using linear interpolation.

(e) Next, an angular histogram similar to (d) is created for each of the spatial bins from (c). Values of the bins are calculated by multiplying the corresponding bin value from (c) by the bin value from (d) and by the length of the projected normal from (b).

(f) The process is repeated for each of the three planes determined by the axes of the reference frame. As a result, the complete descriptor consists of (e) and two other histograms similar to (e).

Descriptors extracted from all nearby points are summed to create the final feature descriptor. The last step is to normalize the descriptor so that the sum of values in all bins is one.

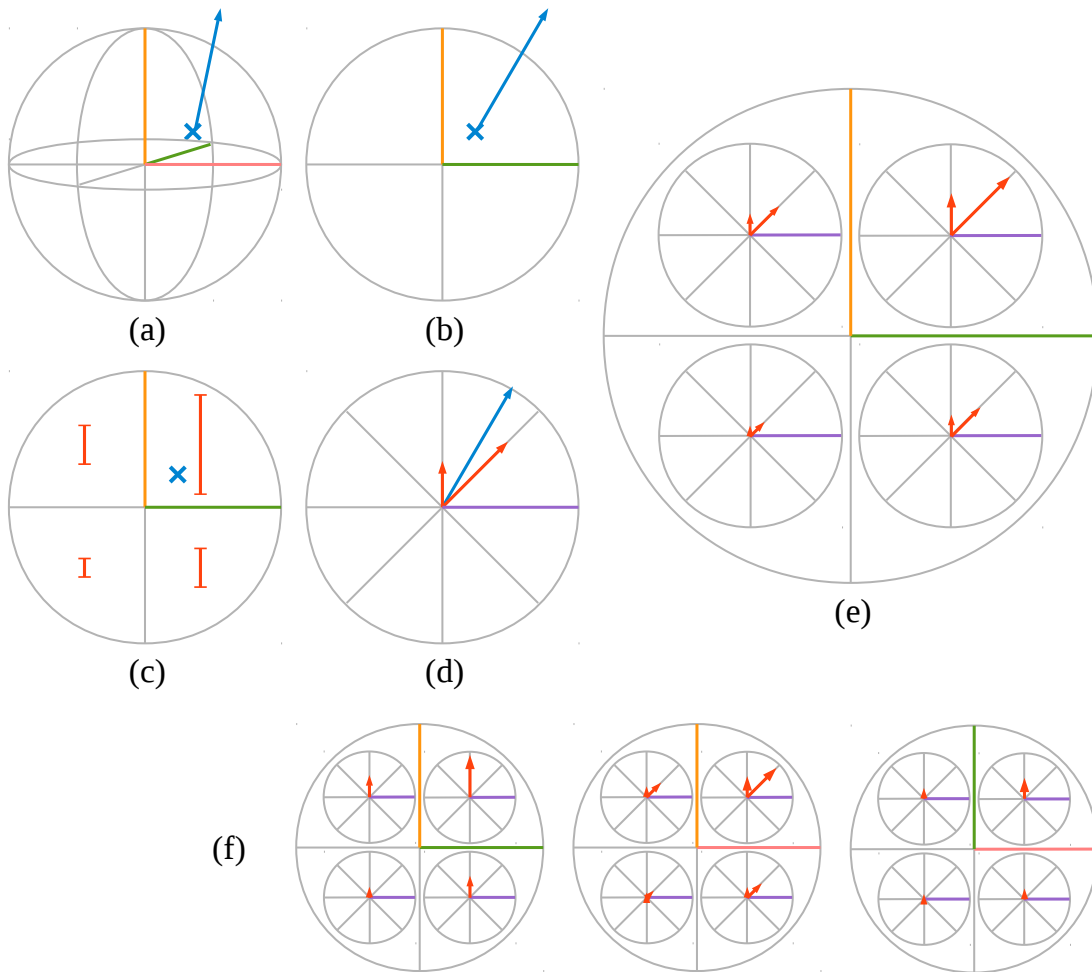


Figure 3.2 Descriptor extraction from a given point with a corresponding surface normal.

3.2.5 Descriptor matching and transformation estimation

To find the estimated relative pose, correspondences – pairs of similar features from the reading and the reference cloud – must be established. This matching process depends on the type of the descriptor; for example, binary descriptors such as BRAND[18] are comparable using the bit-wise exclusive or (XOR) operation. To match the descriptors in the method by Petricek and Svoboda[19], a nearest neighbor search is performed in

a high-dimensional space. Each bin of the histogram is treated as a dimension, and the “distance” (i.e. dissimilarity) between two descriptors is measured by Euclidean distance. Based on the established descriptor pairs, a robust estimator is used to extract the approximate relative transformation. In [19], the RANSAC algorithm is employed.

3.3 Using camera imagery in feature-based registration

3.3.1 Camera projection and 3D gradient direction

In our work, we propose changes to the method by Petricek and Svoboda[19] that make use of visual information available for the NIFTi robot. In particular, we take advantage of camera imagery captured during the scanning process. Since the camera calibration data is available for each video feed, it is possible to project a 3D point into any of the camera images. We implemented a camera projection in MATLAB that complies with the pinhole camera model with two tangential and three radial distortion coefficients[3]. The following computation provides a camera projection (x', y') for a given 3D point (x, y, z) :

$$\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = R \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \mathbf{t} \quad (3.6)$$

$$x_2 = \frac{x_1}{z_1} \quad y_2 = \frac{y_1}{z_1} \quad (3.7)$$

$$r = x_2^2 + y_2^2 \quad (3.8)$$

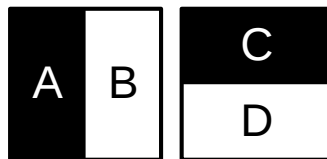
$$\begin{aligned} x_3 &= x_2(1 + k_1r^2 + k_2r^4 + k_3r^6) + 2p_1x_2y_2 + p_2(r^2 + 2x_2^2) \\ y_3 &= y_2(1 + k_1r^2 + k_2r^4 + k_3r^6) + 2p_2x_2y_2 + p_1(r^2 + 2y_2^2) \end{aligned} \quad (3.9)$$

$$x' = f_x x_3 + c_x \quad y' = f_y y_3 + c_y \quad (3.10)$$

where f_x, f_y, c_x, c_y are intrinsic camera parameters, p_1, p_2 are tangential distortion parameters, and k_1, k_2, k_3 are radial distortion parameters. The 3-by-3 rotation matrix R and translation vector \mathbf{t} transform the point from world coordinates into a coordinate system fixed with respect to the camera, whose origin is at the center of the camera projection plane, its Z-axis is the view direction, its Y-axis is vertical and its X-axis is horizontal.

To enable fast extraction of visual information from an image, we make use of integral images[6]. Integral image is a data structure that, once generated for a given image, allows to quickly calculate the sum of values in an arbitrary rectangular area. We use sums of rectangular areas to extract gradient directions in the camera images; this technique is used by Bay et al.[8] to determine feature orientation in SURF. Only the sums of four areas in an image patch are required to compute Haar wavelet responses r_x, r_y :

$$r_x = \sum_{x \in B} I(x) - \sum_{x \in A} I(x), \quad r_y = \sum_{x \in D} I(x) - \sum_{x \in C} I(x) \quad (3.11)$$



The resulting response vector (r_x, r_y) serves as an estimation of the dominant gradient direction in the image patch. With good lighting conditions, the dominant direction is repeatable when viewed from different viewpoints; this motivates the usage of Haar wavelet responses in SURF[8]. We propose to use this technique to estimate a three-dimensional gradient direction for a given point with a specified surface normal. Our method consists of the following steps:

1. First, we project the point in question into the camera images. Since we are considering multiple camera views, we choose the best image for the point, based on the distance of the projection to the nearest edge of the image.
2. In the best image, we consider a square image patch centered at the projection of the point. For this patch, we extract the dominant gradient direction as in equation 3.11. We receive the two-dimensional response vector (r_x, r_y) .
3. Using the 3-by-3 rotation matrix from equation 3.6, we transform the response vector into world coordinates:

$$\mathbf{r}_1 = R^{-1} \begin{bmatrix} r_x \\ r_y \\ 0 \end{bmatrix} \quad (3.12)$$

4. We assume that while the direction of the gradient is oriented as it appears from the viewpoint of the camera, the gradient is also tangential to the surface. Using the surface normal \mathbf{n} , we project the response vector (r_x, r_y) along the camera Z-axis onto the tangent plane of the point:

$$\mathbf{z} = R^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad (3.13)$$

$$\mathbf{r}_2 = (\mathbf{r}_1 \times \mathbf{z}) \times \mathbf{n} \quad (3.14)$$

5. To obtain the final gradient \mathbf{r} , we normalize \mathbf{r}_2 :

$$\mathbf{r} = \frac{\mathbf{r}_2}{\|\mathbf{r}_2\|_2} \quad (3.15)$$

3.3.2 Descriptor extraction

Our goal is to extend the descriptor from section 3.2.4 using the available visual data. We propose to enhance the discriminative power of the descriptor by considering the above mentioned 3D gradient directions. We modify two steps of the algorithm by Petricek and Svoboda[19]:

1. In pre-processing, we add a step to the end of the filtering pipeline that calculates the 3D gradient direction for each point in the point cloud. To achieve scale invariance, we determine the sizes of image patches based on the distance to the camera; for each point (x, y, z) , we also find a point (u, v, w) translated in a direction orthogonal to the direction of the camera:

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} + s \cdot R^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad (3.16)$$

where R is the rotation matrix from (3.6) and s is the intended size of the patch. Both the original and the translated 3D point are projected and their image distance determines the projected size of the patch s' :

$$s' = \|(x', y') - (u', v')\|_2 \quad (3.17)$$

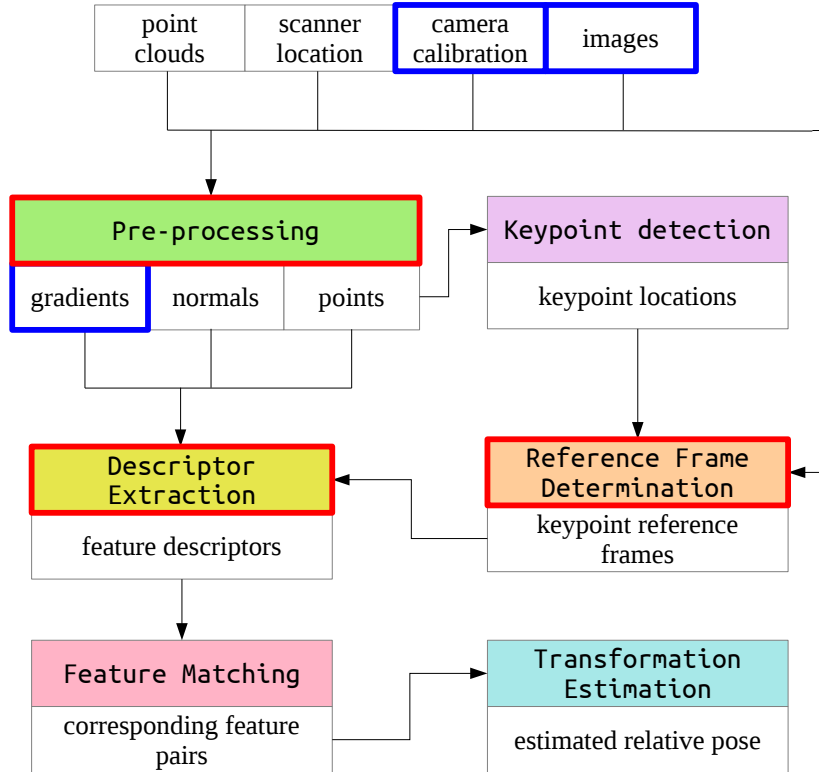


Figure 3.3 Overview of changes to the global method by Petricek and Svoboda[19]. Newly introduced data is marked blue, enhanced steps are marked red.

where (x', y') and (u', v') are camera projections of (x, y, z) and (u, v, w) , respectively. Using the patch and a pre-determined surface normal \mathbf{n} , we follow the equations (3.12, 3.13, 3.14, 3.15) to calculate the 3D gradient direction \mathbf{r} .

2. In descriptor extraction, we extend the descriptor (see section 3.2.4) so that the 3D gradient directions are considered as well as normals. We effectively double the descriptor size, building the original descriptor twice; once for the normals (as previously), and once for the gradient directions.

3.3.3 Reference frame determination

To determine the reference frame, we propose to employ the orientation assignment method from SURF[8]. To briefly describe our intention: in a camera image, we robustly determine the dominant gradient direction near a keypoint. Because the direction is repeatable, we use it as an axis of the keypoint's reference frame. Our changes to the original algorithm from [8] reflect that all results obtained in the image plane (i.e. the camera plane) need to be transformed into world coordinates. Additionally, we are considering a multi-camera system. There are two parameters to our algorithm: radius of samples ρ and image patch size σ . The algorithm constitutes of the following steps:

1. We establish the reference frame $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ from equations 3.4 and 3.5. In the following steps, we use \mathbf{a}_3 as the direction of the surface normal.
2. We improve the robustness of the reference frame determination by repeating the following computation for a set of nearby locations. We choose a set S of 109 evenly spaced positions in the tangent plane of the keypoint; the positions are the centers of perfectly packed circles into a circle with the radius ρ (see figure 3.4).

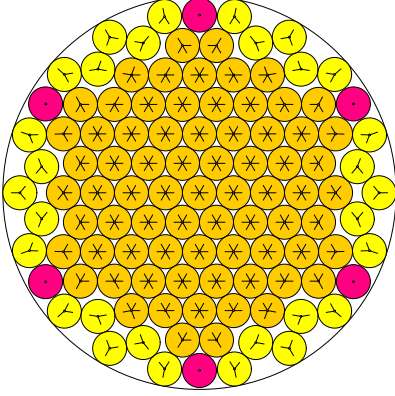


Figure 3.4 The best known packing of 109 equal circles in a circle. Markings in the circles and their colors are irrelevant to our discussion and can be safely ignored. Drawing taken from [2].

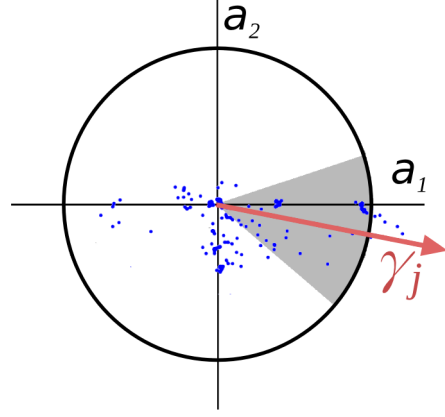


Figure 3.5 Dominant direction determination. In the tangent plane a_1, a_2 , weighted gradients (g'_1, g'_2) (blue) are summed in each sliding window with a fixed size $\frac{1}{3}\pi$. The greatest summed gradient γ_j is the dominant direction. Drawing taken from [8].

3. For each sample position in the set S , we obtain an image patch as in 3.3.2, step 1, substituting σ for the patch size s . For the patch, we compute Haar wavelet responses r_x, r_y (see equation 3.11). We also compute the 3D gradient direction \mathbf{r} using the equations 3.12, 3.13, 3.14, 3.15, with \mathbf{a}_3 as the surface normal \mathbf{n} .
4. We decompose the 3D gradient direction \mathbf{r} into the directions of $\mathbf{a}_1, \mathbf{a}_2$, receiving coordinates (g_1, g_2) in the tangent plane of the keypoint.

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} \mathbf{r}^T \mathbf{a}_1 \\ \mathbf{r}^T \mathbf{a}_2 \end{bmatrix} \quad (3.18)$$

5. We estimate the relative magnitude of the gradient as the magnitude of the response vector (r_x, r_y) divided by the area of the image patch s'^2 . This gives us a scale-invariant result, comparable to results with an arbitrary image patch size. Additionally, we weight the samples with a Gaussian centred at the keypoint, with standard deviation $\frac{1}{2}\rho$. Given the weight w , we obtain the final gradient (g'_1, g'_2) :

$$\begin{bmatrix} g'_1 \\ g'_2 \end{bmatrix} = \frac{w\sqrt{r_x^2 + r_y^2}}{s'^2} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \quad (3.19)$$

6. We calculate the gradient (g'_1, g'_2) for each point in S . In the tangent plane, we select the dominant gradient direction in the same way as in SURF[8]; we compute the sums γ_i of gradients contained within a sliding orientation window i with a fixed size $\frac{1}{3}\pi$. We select the orientation window j with the greatest magnitude $\|\gamma_j\|_2$ (see figure 3.5); then, we conclude that the dominant direction is γ_j .
7. We normalize γ_j and convert it to world coordinates, obtaining the final 3D gradient direction γ . Then, we establish the reference frame $\mathbf{a}'_1, \mathbf{a}'_2, \mathbf{a}'_3$, using γ as one of its axes:

$$\gamma = \frac{1}{\|\gamma_j\|_2} \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 \end{bmatrix} \gamma_j \quad (3.20)$$

$$\mathbf{a}'_2 = \gamma \quad \mathbf{a}'_3 = \mathbf{a}_3 \quad \mathbf{a}'_1 = \mathbf{a}'_2 \times \mathbf{a}'_3 \quad (3.21)$$



Figure 3.6 Unit gradient directions \mathbf{r} at sample points (green), weighted by their distance from the keypoint, and the resulting dominant direction γ (red). Gradient magnitudes used to determine γ are not visualized in this drawing.

3.4 Dataset

To test our novel contributions, a dataset is needed that provides camera imagery. Petricek and Svoboda[19] test their method on the challenging datasets[22] that we use to compare local methods (see section 2.3). However, these datasets contain no visual information, making it impossible to test our proposed method.

To create our own dataset, we have used data captured by the NIFTi robot. This provides us with 6 video feeds from the PointGrey Ladybug 3 omnidirectional camera, with 2 Mpx resolution each, and range data from the Sick LMS-151 laser rangefinder, with point clouds ranging in number from 35.000 to 50.000 points. The point clouds are aligned by the iterative closest point method by Kubelka et al.[15]; we use the result of ICP registration as ground truth.

We have extracted a series of 176 point clouds along with the corresponding camera images from a recording of a courtyard at the CTU campus at Charles Square. During the scanning process, the robot moves in an outdoor, building-surrounded environment. To select point cloud pairs, we followed the procedure described in [21]: for each pair of the clouds, we have estimated the overlap ratio (see figure 3.8), and based on that information we have selected 35 pairs of scans, with overlap ratios distributed uniformly between 30 % and 99 %.

We do not consider any perturbations of the initial pose in this dataset. There are two reasons to do so: firstly, global registration methods provide similar results for arbitrary initial poses, therefore perturbations are unnecessary to test these methods. Secondly, our cameras are calibrated so that only the point clouds in the undistorted initial pose are projected correctly into the camera images.

3.5 Experimental results

Although our dataset is small and its ground truth is relatively imprecise, it is sufficient for our purposes. We compare the original, unaltered method by Petricek and Svoboda[19] to the method that contains one of the proposed changes. At the same time, we take the opportunity to test various parameter configurations of our version



Figure 3.7 A set of camera images providing visual information for one of the point clouds in our dataset.

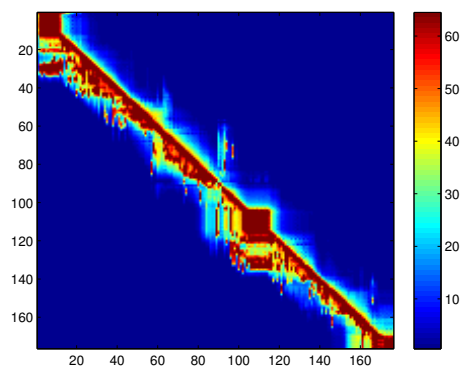


Figure 3.8 Estimated overlap ratios of point clouds in our dataset. Overlap ratio is the ratio of surfaces in point cloud A that are also present in point cloud B. Dark red color signifies near maximum overlap; this is the case for pairs of point clouds that were captured consecutively. Dark blue color marks no overlap between the two point clouds.

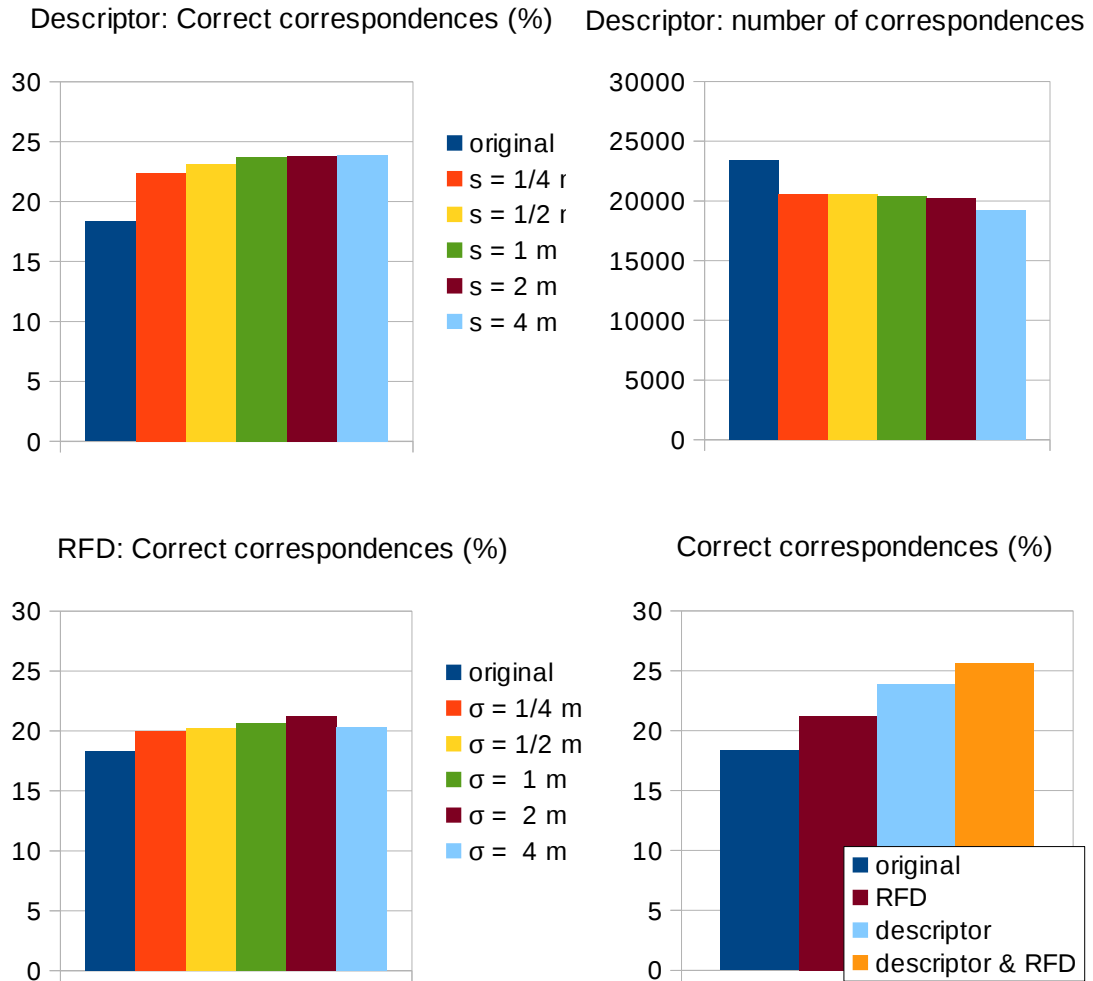


Figure 3.9 Results of testing our method compared to the original, unaltered one. Top left, bottom left: All tested configurations of our changes show an increase of the ratio of correct correspondences, suggesting increased precision. Top right: all descriptor configurations also decrease the total number of correspondences, indicating increased discriminative power. Bottom right: a comparison of the best configurations of the descriptor and the reference frame disambiguation (RFD).

of the method.

To compare the precision of the methods, we consider correspondences, i.e. pairs of corresponding features extracted from the clouds by the matching process (see section 3.2.5). From these pairs, we count those that are located at the same position after the ground truth transformation is applied to the reading cloud – these are the correct correspondences. Figure 3.9 explores the ratio of the number of correct correspondences to the total number of correspondences over the whole dataset. An increased ratio of correct correspondences is an indication of greater precision of a method. Results are shown separately for the descriptor and the reference frame determination; in both cases, the ratio of correct correspondences has increased significantly, suggesting that our novel contributions are indeed an improvement in terms of precision.

For descriptor configurations, the total number of correspondences over the whole dataset is also shown in figure 3.9, demonstrating a decrease; this indicates that the matcher has refused some pairs due to their dissimilarity, suggesting that the discrimi-

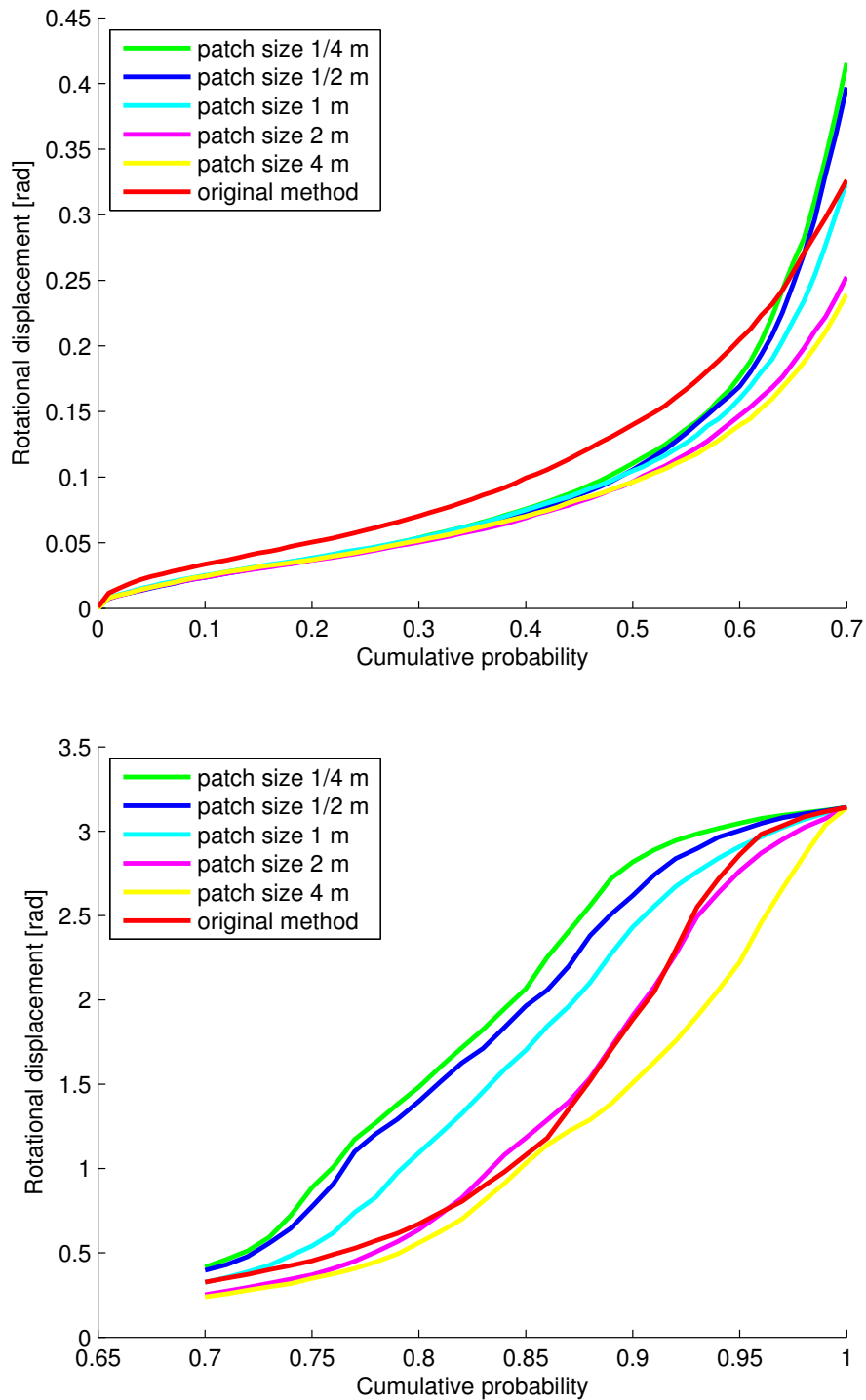


Figure 3.10 Results of testing our reference frame determination method for various values of image patch size σ , with sampling radius ρ set to 1 m. The graphs show quantiles of reference frame rotation error. All variations of our method beat the original method for any quantile up to 0.6, showing an increase of precision. Beyond 0.6-quantile, it is suggested that the original method is more robust than most of our configurations, however our method with patch size $\sigma = 4$ m is superior.

native power of the descriptor has increased.

Additionally, regarding reference frame determination, we consider rotation displacement of the reference frames. First, we identify the keypoints that have the same position in the clouds after being aligned using ground truth. For each pair, we compute the displacement of the reference frames using equation (2.5). We gather the rotation error of reference frames over the whole dataset, and analyze it using quantiles.

Figure 3.10 shows the quantiles of the rotation error. Our method excels in the 0 to 0.6-quantile range, suggesting its precision is an improvement over the original method in most cases. Over the 0.6-quantile, the original method beats most of our configurations, showing that it is quite robust. However, one of our configurations ($\sigma = 4$ m) overcomes all other tested methods for any quantile.

3.6 Conclusion

We have proposed, implemented and tested two independent enhancements to the method by Petricek and Svoboda[19]: a new descriptor, and a new method of reference frame determination. Both of our contributions are a success, showing significantly improved capabilities over the original, unaltered method. We have achieved the goal of improving the original method based on the availability of visual information, in particular camera imagery.

We have successfully applied solutions from image registration methods, specifically SURF[8], to point cloud registration. As the challenging datasets by Pomerleau et al.[22] lack visual information, we have created our own dataset, based on range and visual data captured in an outdoor environment. The result of our work is a competitive global registration method. With that said, we do not believe that the possibilities of color-aware point cloud registration are exhausted. On the contrary, the subject matter is still largely unexplored, creating room for future work.

4 Conclusions

In our work, we focused on point cloud registration methods. A different approach was employed for each of the two classes of methods, local and global; local registration methods were approached investigatively, offering an insightful view into their capabilities. For the global methods however, a more generative approach was applied, creating a method that takes advantage of visual data.

For the purposes of local method comparison, experiments were carried out that make use of an existing, publicly available protocol, based on a number of high-quality datasets. To inspect the capabilities of the methods in a greater detail, an additional protocol based on the same datasets was created; by evaluating our protocol, it is possible to study the capabilities of the methods in an unprecedented detail, overcoming any previous work that we know of. Limitations of the examined methods are revealed in the form of maximum viable initial pose displacement, provided that error requirements are given. Additionally, results for methods that were neglected in our experiments can be received at a later date and directly compared to ours.

Analyzing the results, it is shown that of the tested methods, the iterative closest point (ICP) algorithm configured by Kubelka et al.[15], and the three-dimensional normal distribution transform (3D-NDT) algorithm implemented in the Point Cloud Library[5] share the lead in registration quality. Using these methods for point cloud registration is recommended. Two composite methods were created and tested, but were not proven useful.

Concerning global methods, a feature-based method by Petricek and Svoboda[19] was enhanced using visual information from cameras. Two changes have been proposed, an extension of the descriptor, and a modification of reference frame determination. Parts of the SURF[8] algorithm have been used to extract visual information, introducing an image registration technique into point cloud registration. A dataset containing visual data was created to test our proposals, along with a testing protocol.

Using the protocol, the original method was evaluated, as well as its modifications. The modifications are shown to be effective, overcoming the unaltered version of the method; extended descriptor increases the number of correct feature correspondences, and changes in reference frame determination decrease the rotation error of the established reference frames. The goal of improving a feature-based registration method by fusing visual and range data was accomplished.

We believe that the subject of visual data-enhanced point cloud registration is not yet fully explored. Our suggestions of future work include: exploring the use of colored point clouds, as opposed to camera imagery; investigating three-dimensional binary descriptors, e.g. a modification of BRAND[18], in the context of visual data; using visual data to improve the saliency measure for keypoint detection; and extending our method to make use of color information other than intensity.

Bibliography

- [1] Adaptive traversability. http://cw.felk.cvut.cz/wiki/misc/projects/nifti/sw/adaptive_traversability. Accessed on May 23, 2014. 3
- [2] The best known packings of equal circles in a circle. <http://hydra.nat.uni-magdeburg.de/packing/cci/cci.html>. Accessed on May 23, 2014. 28
- [3] Camera calibration and 3D reconstruction. http://docs.opencv.org/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html. Accessed on May 23, 2014. 25
- [4] NIFTi human-robot team returns from earthquake deployment in Italy. <http://vision-robotics.blogspot.cz/2012/08/nifti-returns-from-earthquake.html>. Accessed on May 23, 2014. 3
- [5] PCL - point cloud library. <http://pointclouds.org/>. Accessed on May 23, 2014. 10, 34
- [6] Summed area table. http://en.wikipedia.org/wiki/Summed_area_table. Accessed on May 23, 2014. 25
- [7] Michel A. Audette, Frank P. Ferrie, and Terry M. Peters. An algorithmic overview of surface registration techniques for medical imaging. *Medical Image Analysis*, 4(3):201 – 217, 2000. 2
- [8] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008. 22, 25, 26, 27, 28, 33, 34
- [9] P.J. Besl and Neil D. McKay. A method for registration of 3-D shapes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 14(2):239–256, Feb 1992. 4, 10
- [10] P. Biber and W. Straßer. The normal distributions transform: a new approach to laser scan matching. In *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 3, pages 2743–2748 vol.3, Oct 2003. 4, 6
- [11] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. In *Robotics and Automation, 1991. Proceedings., 1991 IEEE International Conference on*, pages 2724–2729 vol.3, Apr 1991. 4, 10
- [12] Jan Elseberg, Stéphane Magnenat Rol, and Siegwart Andreas Nüchter. Comparison of nearest-neighbor-search strategies and implementations for efficient shape registration, 2012. 4
- [13] B. Huhle, Martin Magnusson, W. Strasser, and A.J. Lilienthal. Registration of colored 3D point clouds with a kernel-based extension to the normal distributions transform. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 4025–4030, May 2008. 6

- [14] A.E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(5):433–449, May 1999. 21, 22, 23
- [15] Vladimír Kubelka, Lorenz Oswald, François Pomerleau, Francis Colas, Tomáš Svoboda, and Michal Reinstein. Robust data fusion of multi-modal sensory information for mobile robots. *Journal of Field Robotics*, in press. 4, 5, 9, 10, 20, 29, 34
- [16] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004. 22
- [17] Martin Magnusson. *The Three-Dimensional Normal-Distributions Transform — an Efficient Representation for Registration, Surface Analysis, and Loop Detection*. PhD thesis, Örebro University, December 2009. Örebro Studies in Technology 36. 4
- [18] E.R. Nascimento, G.L. Oliveira, M. F M Campos, A.W. Vieira, and W.R. Schwartz. BRAND: A robust appearance and depth descriptor for RGB-D images. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 1720–1726, Oct 2012. 22, 23, 24, 34
- [19] T. Petricek and T. Svoboda. Point cloud registration from local feature correspondences - evaluation on challenging datasets. Unpublished work, under review. 22, 23, 24, 25, 26, 27, 29, 33, 34
- [20] T. Petricek and T. Svoboda. Area-weighted surface normals for 3D object recognition. In *ICPR'12*, pages 1492–1496, 2012. 22, 23
- [21] François Pomerleau, Francis Colas, Roland Siegwart, and Stéphane Magnenat. Comparing ICP variants on real-world data sets. *Autonomous Robots*, 34(3):133–148, 2013. 4, 5, 8, 10, 29
- [22] François Pomerleau, Ming Liu, Francis Colas, and Roland Siegwart. Challenging data sets for point cloud registration algorithms. *The International Journal of Robotics Research*, 31(14):1705–1711, 2012. 7, 8, 29, 33
- [23] R.B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (FPFH) for 3D registration. In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 3212–3217, May 2009. 21, 22, 23
- [24] T. Stoyanov, Martin Magnusson, and A.J. Lilienthal. Point set registration through minimization of the L2 distance between 3D-NDT models. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 5196–5201, May 2012. 6
- [25] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *11th European Conference on Computer Vision (ECCV 10)*, pages 356–369, Hersonissos, Crete, Greece, September 2010. 21, 22, 23