Ground Truth Free Evaluation of Segment Based Maps

Rolf Lakaemper

***Abstract*— For ground truth free map evaluation, map intrinsic properties have to be utilized in order to define a measure of quality. This paper suggests the property of map consistency, based on the similarity of different model elements describing a single feature in the environment. The evaluation algorithm is tailored for segment based maps. Respecting the segment representation, it inherits the properties of memory- and runtime efficiency. In comparison to classic grid based evaluation models, which are mostly used for evaluation of maps represented by points, it is also able to detect structural inconsistencies of maps, even if the model elements (the segments) overlap. The paper explains the core technique, hierarchical clustering of segments, illustrates properties of the underlying segment similarity measures and cluster confidence measure. Experiments on different segment based maps show its consistency with visual inspection.**

I. Introduction and Approach

Robot maps are the basis for nearly all autonomous navigational tasks in robotics. It is therefore an important task to evaluate the performance of robot mapping systems with respect to different parameters, e.g. precision, completeness and consistency. Mapping, in general, is spatial analysis of environmental features of interest. Inherent to this process is its task dependency, hence there is no 'optimal general mapping'. As a direct consequence, there is no single best strategy to evaluate maps. Completeness might be preferred to precision in path planning tasks, while precision might be of higher importance to locally bounded tasks. In fact, map evaluation turns out to be a field with many different aspects. The showcases of precision and completeness are parameters which can only be evaluated if a trusted model, the ground truth, is available.

The field of robot map evaluation opens up even wider if ground truth independent techniques are taken into account as well. The parameter of consistency, i.e. a parameter that describes a map intrinsic property, replaces the aforementioned examples in such cases: precision and completeness can only be estimated based on the consistency of features.

Both, ground truth based and non ground truth based evaluation deal with aspects of distance; either distance to an external reference, or an intra-distance between elements of a single map. The distance aspect is expressed by the choice of a particular distance-measure (which is not necessarily distance *metric*). The most common choices, Euclidean distance of map-features or angular/translational (pose) distance, lead to *grid-based* evaluation, and *pose-based* evaluation respectively.

Rolf Lakaemper is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA lakamper@temple.edu

Additionally, design of map a evaluation strategy is an engineering process, in the sense that it should respect advantages/disadvantages of different data representations. In practice, the way of robot internal representation of the physical world is carefully chosen; not following the specific data representation, but transforming the model, might render the evaluation useless (an example is given in section VII-A, where map evaluation is suggested to provide a break condition for an iterative mapping process).

An example for such a case is the evaluation described in this paper: it respects the underlying data structure of line-segments. One argument for line-segment based maps is its efficiency. A transformation to point based map, e.g. for grid based evaluation, could be too costly in cases where online, real time evaluation is needed. Using the aforementioned categories, the presented evaluation

- is not ground truth based
- measures the map consistency
- works on line segments as underlying data structure

and

- is neither grid, nor pose based, but uses a segment distance measure, which models the dissimilarity in visual perception of pairs of line segments.

Intuitively, the presented method combines properties of both, pose and grid based methods: as in grid based techniques, it infers consistency from feature density. As in pose based techniques, the density computation includes angular differences. However, it relates closer to grid based techniques, since only a single (merged) map is taken as input; pose based approaches need global maps consisting of multiple single scans. Since single line segments, not single scans (= sets of segments) are compared, it is, like grid based evaluation, of non-rigid nature.

The presented method evaluates the consistency of segment based robot maps. Compared to raw points, e.g. gained from laser range scanners, line segments can be seen as higher geometric structures. They contain, additional to the pure location information, a direction, and can also be seen as labels. A single line segment subsumes points belonging to a single linear structure in the environment. The interest in robot mapping based on higher geometric structures like line segments is currently growing. Obvious advantages in runtime, memory efficiency and simpler mid-level analysis capability make such mapping approaches powerful competitors to the classic, point based techniques.

An important design paradigm of the presented research is not to leave the very efficient and compact data representation by segments. Such an approach leads to multiple

advantages compared to point/grid based methods:

- The segment based approach is fast. In indoor environments or urban outdoor environments, a typical scan consists of $n < 20$ segments of sufficient length, while the number of data points is typically one to two orders of magnitude (factor $10-100$) higher. This becomes especially important when point relations between different scans have to be evaluated, which usually implies algorithms with runtime between $O(nlogn)$ and $O(n^2)$.
- The segment based approach is memory efficient. Compared to occupancy grids, the memory consumption is significantly lower.
- The segment based approach is precise. Segment endpoints don't have to be adjusted to a resolution parameter, hence there are no quantization errors. This is in contrast to grid based approaches.
- The segment based approach captures structural information. Even if line segments from different single scans are united to a single set, the global map, they still represent points belonging to a single structure, and they still have their own direction property. If two segments, which originate from different scans and represent different features, are erroneously transformed to be at the same location in the global map, they can still be distinguished if their direction differs. Hence, this information goes significantly further than the information of pure object presence (location), contained in raw point data. Figure 6 to 8 show examples: the red segments are detected as noise. This example shows a clear advantage of line based consistency evaluation to its point based counterpart: in point density based approaches, the original data points would not only be seen as correct (since the point density in this area is high), but would have even enhanced the confidence in this region, since the density of points is high. Measuring the similarity not only by Euclidean distance, but taking directional properties into account, their distance is higher, which correctly leads to a lower density.

The basic idea of the approach is to cluster segments, based on an inter segment-distance measure. The quality of clusters defines the confidence in the participating segments, which in turn defines the confidence in the entire map. The main steps are i) the definition of a perceptually consistent segment-distance measure, ii) the adaption of a classic clustering technique (hierarchical clustering) to gain a parameter free clustering system, and iii) a new measure for intra cluster consistency, which directly leads to the final goal, the confidence measure. This paper illustrates these steps, for technical details please see [7].

II. Related Work

To the authors's best knowledge, there are no publications available about generic map evaluation, the reason being that map evaluation is highly task specific. Task specific map evaluation is usually performed in the broader environment of robot competitions, such as RoboCup [1] or the US Department of Energy Grand Challenge [13]. Test arenas, developed by the National Institute of Science and Technology (NIST) exist [3], as an effort to create robot maps in standard environments. These arenas were used in various events, e.g. the RoboCup Rescue competition and the Response Robot Evaluation Exercise [11].

An occupancy grid based evaluation tool, the Jacobs Map Evaluation Toolkit [2], was utilized in the RobocupRescue competition 2008. Aside from functionalities like ground truth map creation, it consists in its core of a metric comparing the (grid/pixel based) maps. In short, correspondences between foreground points of the evaluated map and a ground truth map are established. The correspondence quality is computed using the spatial distance of the corresponding points.

In contrast, the presented evaluation method does not perform a comparison to a ground truth map, but aims to analyze the consistency of a single map. Working on a higher data structure, line segments, it tries to capture regional structural properties. These are evaluated based on their ambiguity of representation: a single cluster represents a single feature, a high intra cluster distance can be interpreted as ambiguity, or low confidence.

A more general introduction and overview of benchmarking and evaluation in robotics is given in [10].

III. Distance Measure for Segment Pairs

This section introduces a distance measure between pairs of line segments s_1, s_2. The basic idea of the distance measure is to merge two line segments to an 'average' segment $\bar{s}$. The distance is the merging cost, which consists of three parts:

- the angular distance between s_i and $\bar{s}$, $i = 1,2$
- the spatial distance between s_i and $\bar{s}$, $i = 1,2$
- the spatial distance between s_1 and s_2.

The first two parts penalize the amount of 'non collinearity' of the segments, the third part penalizes spatial distance. Although used as a distance measure between two segments, the design is based on comparison to a 'virtual' average segment. This is motivated by certain experiments, suggesting that human perception assigns or connects line segments to larger structures under certain circumstances. The definition of the measure is out of scope of this paper and will be part of a future publication. Figures 1 to 4 illustrate the properties of the similarity measure.

Figure 5 gives examples of random segment configurations and resulting distances.

IV. Clustering

For the clustering, agglomerative hierarchical clustering in 'single' mode is utilized. This method seeks to build a bottom up hierarchy of clusters, starting with each segment being a single cluster, ending in a single cluster containing all segments. Pairs of clusters are merged as one moves up the hierarchy. The merge is determined in a greedy manner: the two clusters with minimal distance are merged to a single one. Hierarchical clustering allows for different strategies to determine the distance of the newly emerged cluster to the remaining elements. In our case, we use the

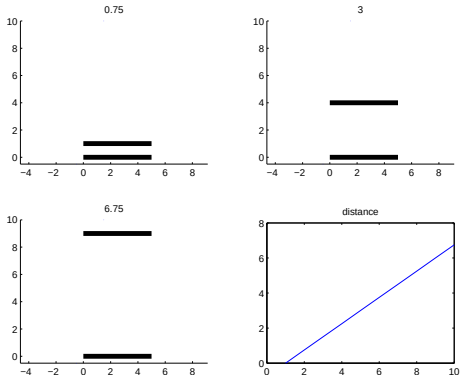

Fig. 1. The distance measure increases linearly with distance (if segments are parallel).

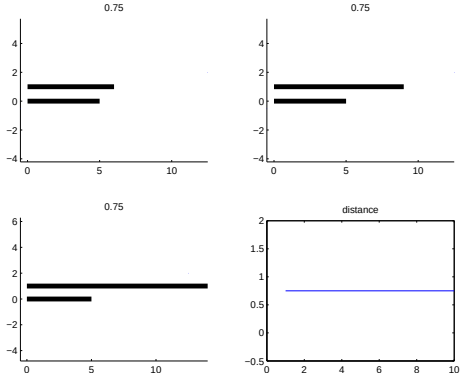

Fig. 2. The distance measure is constant when a single segment is elongated (if segments are parallel). This is an important property for laser scans: scanned from different positions, identical linear structures can be represented by segments of different length. Compare to Figure 4.

'single mode' strategy: the distance between two clusters is the minimum distance between their elements. There is a geometric motivation for the use of this mode: in the example of collinear, slightly overlapping segments single mode clusters these segments to a single group — intuitively, single mode clustering acts like a connected components algorithm, the necessary topology being defined through the distance measure (small distance = neighbors).

Hierarchical clustering has two main properties which suggest its use in the segment merging context: first, it is, in its first stage, parameter free, i.e. no pre-defined number of clusters has to be determined. Parameters might be introduced later in a follow up stage, which selects the level of clustering (agglomerative hierarchical clustering always ends in a single cluster). Second: it is simply based on mutual distances between the data points (here: line segments), yet without the need to embed them in a metric space. This means, hierarchical clustering can deal with any distance measure (especially non-metrics, as in the given case). More details about hierarchical clustering in general can be found e.g. in [4]. See [7] for technical details.

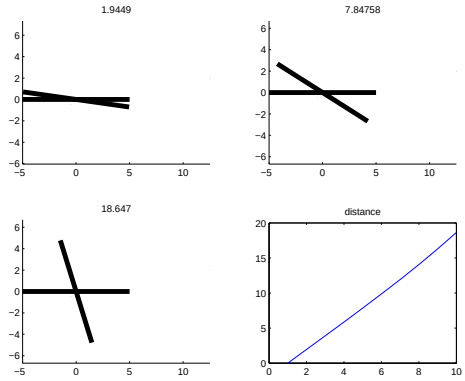

Fig. 3. The distance measure increases non-linear with rotation.

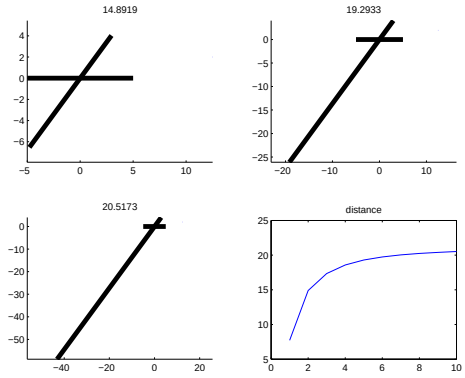

Fig. 4. The distance measure increases non-linear when a single segment is elongate, non-parallel case. Compare to Figure 2.

V. The Core of the Evaluation: Measuring the Cluster Quality

The main step in our evaluation is to determine the consistency of each cluster. We introduce a new intra-cluster-consistency measure $\mathscr{C}$ which to the specific problem of segment cluster evaluation.

In $\mathscr{C}$, collinear structures are favored, while clusters containing wide-spread segment sets are penalized. Similar to the segment distance measure, each segment in the cluster is compared to an average cluster segment, the cluster representative. In analogy to classic intra cluster consistency measures, the angular and spatial distance to this representative is taken into account to determine the cluster consistency. Intuitively, all angular distances of segments in one cluster to the average cluster representative are computed, as well as the transitional distances. For angular and translational distances, two separate confidence measures $\mathscr{C}_a, \mathscr{C}_t \in [0..1]$ (angular/translational respectively) are computed. The final confidence $\mathscr{C}$ is computed as

$$\mathscr{C} = \min(\mathscr{C}_a, \mathscr{C}_t). \tag{1}$$

A high confidence (close to 1) is only assigned if both, angular and translational confidence are high. Additionally, clusters must contain a certain minimal number of segments (in the current system: three segments), otherwise they are

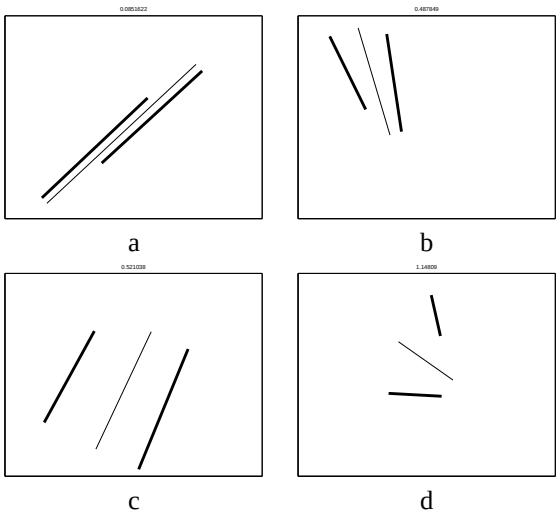

Fig. 5. Segment configurations with increasing distance. a) 0.09 b) 0.49 c) 0.52 d) 1.14. The thin line is the merged segment. The increase in a)-c) results from larger intra segment distance, while d) results from angular distance.

assigned a confidence of $\mathscr{C} = 0$. For details about $\mathscr{C}_a$ and $\mathscr{C}_t$ please see [7]. Figures 6 to 8 illustrate properties of the confidence measure.

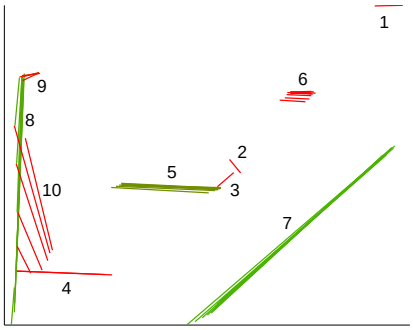

Fig. 6. Evaluating clusters using the confidence measure $\mathscr{C}$. The red/greenness is determined by confidence (the greener, the more confident). Please see text for further explanations.

Figure 6 shows a part of a global map, consisting of 10 aligned scans. This part shows noise and inconsistencies due to various sources. Cluster 1, 2, 3 and 4 are evaluated as inconsistent, since they are single segments, only detected once. They could be correctly sensed objects, only seen once due to the robot's changing position, or dynamic objects (e.g. walking people), sensor noise or objects only seen under a certain (unwanted) tilt of the robot. The segments of cluster 6 and 10 result from the robot scanning the ground while standing on a tilted platform. This kind of noise can be hard to distinguish from non-ground, 'real' objects, since it usually leads to long segment. However, the nature of such noise is that it leads to wide-spread clusters of parallel, non-collinear lines, which are detected by the inconsistency measure. The segments of cluster 9 result in a low consistency, since, given their small length, their locations and directions lead to a high dissimilarity. Only the significant and consistent features represented by clusters 5, 7 and 8 are evaluated as consistent, which is in accord with visual inspection. Figure 7 is an

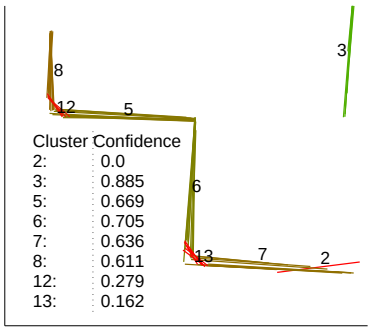

Fig. 7. Evaluating clusters using the confidence measure $\mathscr{C}$. The red/greenness is determined by confidence (the greener, the more confident). Please see text for further explanations.

interesting example which shows the advantages of segment based evaluation over point based systems. Clusters 3, 5, 6, 7 and 8 correctly yield a high confidence measure.Clusters 12 and 13 are fragments from scanned ground, while the robot was tilted, similar to cluster 10 in Figure 6. However, here they overlap with the correct corners, which makes them undistinguishable in a point based system. Only due to their different direction, captured in the segment data structure, they are detectable and can be singled out as inconsistency. In cluster 2 a single noisy segment is detected, which also overlaps with the correct data in cluster 7.

Fig. 8. Evaluating clusters using the confidence measure $\mathscr{C}$. The red/greenness is determined by confidence (the greener, the more confident). right: a magnified view of the marked part of left figure. See text for further explanations.

Figure 8 shows an example of correctly scanned features, which are slightly misaligned. Again, due to the handling of different directions in the consistency measure and segment distance, the lower map quality (compared to correct alignment) can be detected. Detection of such areas is not possible with occupancy grids, but only with methods detecting underlying structural information.

VI. From Regional to Global Evaluation

It is a small step from regional evaluation of single clusters to global map evaluation. Given all clusters C_i along with their confidence measure $\mathscr{C}(C_i)$, we define the global confidence $\mathscr{M}$ of a map by

$$\mathscr{M} = \frac{\sum_i \#C_i \, \mathscr{C}(C_i)}{\sum_i \#C_i} \tag{2}$$

with $\#C_i$ denoting the cardinality of C_i. $\mathscr{M}$ computes the average consistency of all segments, defining the confidence of a segment by the consistency of the cluster it participates in.

VII. Results

A. Evaluation of a Mapping Algorithm During its Performance

This experiment evaluates the result of an interactive alignment algorithm (FFS with Virtual Scans, [6]) after each iteration. Such an evaluation can be used to determine an appropriate stop case for the alignment. The data set 'NIST' consists of 64 scans; in each iteration a new pose for each single scan is computed, using a dynamic- step width gradient descent strategy. Figure 9 shows different iterations. Figure 10 shows the result of the evaluation function applied to each iteration result, and the potential of the resulting map (see [9]). The potential is the target function which FFS optimizes, it is also used as the stop-criterion of the algorithm. Intuitively it describes the visual fitness, or map quality (the lower the better). In both functions, the increasing map quality can be detected (the jitter in both functions comes from the dynamic step-width, which makes FFS over-shoot in the initial iterations to escape local minima). However, the potential detects a significant difference between iteration 100 and 174, which is not in accord with visual inspection (see 9(c) and (d)). The reason is, that the potential is influenced by certain parameters which converge in later iterations (>100), it is therefore not able to detect the (intuitively) visual convergence in earlier stages. The segment based evaluation $\mathscr{M}$ (Figure 10,(a)) is strongly in accord with visual inspection, showing convergence of the alignment in earlier iterations.

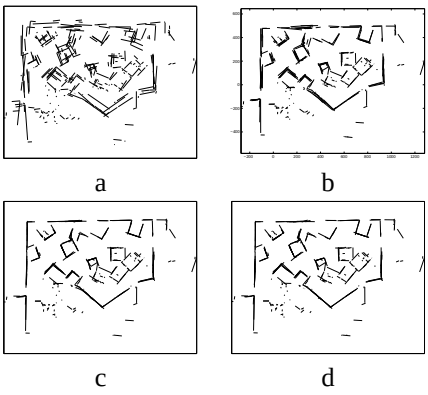

Fig. 9. Iterations 1(a), 35(b), 100(c) and 174(d) of FFS with Virtual Scans on data set NIST.

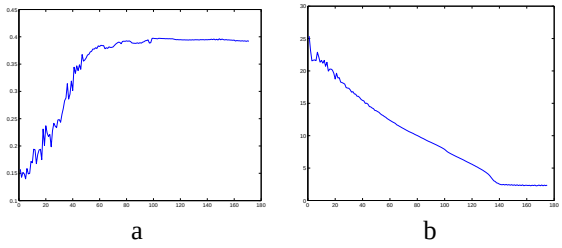

Fig. 10. (a) Result of evaluation during alignment of data set NIST (see also Figure 9). (b) Potential-function of FFS. X-axis: iteration index (1-174). Y-axis: $\mathscr{M}$ (a) and potential (b).

B. Comparative Evaluation of Mapping Algorithms

This experiment compares results of three different mapping algorithms. The first algorithm [5] is a point based alignment (not segment based) algorithm. However, it results in corrected poses of single scans. We applied an algorithm explained in [8] to extract segments from these single scans, and superimposed them, transformed by the corrected poses (Figure 11,a). The second map (Figure 11,b) was computed by a new, segment based algorithm, which we will describe in a future publication. The third map, Figure 12, was created using FFS [9], yet with deliberately mistuned parameters (the parameter σ, which controls the influence of neighboring structures (see [9]), was set to an unreasonably large value). It results in a map which does not represent the environment correctly, *yet contains locally consistent structures*. All output maps consist of the same segments single scans. They differ in the alignment, resulting (as poses) from the different algorithms.

It can be seen that the first map is less consistent than the second. Our evaluation algorithm does not only capture the overall difference in quality (quality of first map: $\mathscr{M} = 0.2769$, quality of second map: $\mathscr{M} = 0.4355$), but also identifies the higher confidence of certain regions, indicated by the color green in Figure 11.

The evaluation of the third map (Figure 12) yields a value of $\mathscr{M} = 0.3558$, and therefore falsely suggests the map to be of higher quality than map one (11,(a)). The reason is, that the algorithm measures consistency, but, lacking a ground truth, can not reliably estimate precision and completeness. A ground truth free evaluation always works under the assumption that the input map models the environment (intuitively) close enough to the reality, i.e. with sufficient precision and completeness, which is not given here. An even more extreme example would be, if the input map would contain single scans without any overlap, i.e. the alignment algorithm locates the scans far apart from each other. Such a global map is entirely consistent, although it does not model reality at all.

The result of this experiment: under the assumption that the input-map models the environment close enough to reality, the algorithm evaluates the map in conformity with visual inspection.

Fig. 11. Evaluation of two maps of the data set 'Freiburg082'. a) Quality of map $\mathscr{M} = 0.2769$. b) $\mathscr{M} = 0.4355$. Colors: level of green (vs red) shows confidence: The greener, the more confident, the more red, the worse. The higher regional confidence in (b) leads to the better total confidence value.

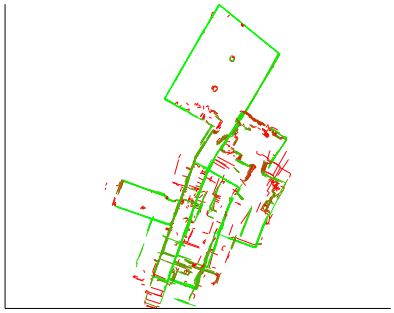

Fig. 12. A map of the same data set as Figure 11, created with deliberately mistuned parameters. Although, by visual inspection, it obviously models the reality worse than Figure 11,(a), its evaluation value is higher due to its higher consistency. Please see text for explanations.

VIII. Runtime

The presented algorithm has an order of magnitude of $O(n^2)$, n = total number of segments, which results from computation of the pairwise segment distance matrix. The MATLAB implementation of the algorithm needed 1 second for the experiment using data set NIST (332 segments), and 5 seconds for the experiment using data set Freiburg082 (1975 segments), both on a 1.8GHz laptop PC.

IX. Conclusion and Outlook

The presented confidence measure evaluates maps in consistency with visual perception. In its core, it uses a classical clustering algorithm, hierarchical clustering, which is adapted to the current problem utilizing a segment distance measure and a segment based cluster confidence measure. Since segment based representation captures structural features better than its lower representation counterpart, point based maps, erroneously mapped/aligned features can be detected even if they overlap with correct features. This leads to detection of structural consistency, which is the main property evaluated by the presented approach. With a redefinition of segment distance and cluster confidence, the approach is extendable to 3D, which makes it interesting for 3D mapping algorithms based on planar elements, e.g. [12].

X. Acknowledgements

Thanks to Alexander Kleiner, University of Freiburg, for the data set 'Freiburg082', and to Jan Elseberg, Jacobs University, Bremen, for providing the maps of Figure 11.

References

[1] S. Balakirsky, S. Carpin, A. Kleiner, M. Lewis, A. Visser, J. Wang, and V. A. Ziparo. Towards heterogeneous robot teams for disaster mitigation: Results and performance metrics from robocup rescue. *Journal of Field Robotics*, 24(11-12):943–967, 2007.

[2] I. Varsadan, A. Birk, and M. Pfingsthorn. Determining Map Quality through an Image Similarity Metric. In *Proceedings of the RoboCup Symposium*, July 2008.

[3] A. Jacoff, E. Messina, B. Weiss, S. Tadokoro, and Y. Nakagawa. Test arenas and performance metrics for urban search and rescue robots. In *Intelligent Robots and Systems, 2003. (IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 4, pages 3396–3403 vol.3, Oct. 2003.

[4] S. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, September 1967.

[5] A. Kleiner and C. Dornhege. Real-time localization and elevation mapping within urban search and rescue scenarios: Field reports. *J. Field Robot.*, 24(8-9):723–745, 2007.

[6] R. Lakaemper. Improving sparse laser scan alignment with virtual scans. In *International Conference on Intelligent Robots and Systems (IROS08)*, Nice, France, September 2008. IEEE.

[7] R. Lakaemper. A confidence measure for segment based maps. In *Performance Metrics for Intelligent Systems, PerMIS 2009*, Gaithersburg, MD, September 2009. NIST.

[8] R. Lakaemper. Simultaneous multi-line-segment merging for robot mapping using mean shift clustering. In *International Conference on Intelligent Robots and Systems (IROS09)*, St Louis, MO, USA, September 2009. IEEE.

[9] R. Lakaemper, N. Adluru, L. Jan Latecki, and R. Madhavan. Multi robot mapping using force field simulation: Research articles. *J. Field Robot.*, 24(8-9):747–762, 2007.

[10] R. Madhavan, R. Lakaemper, and T. Kalmar-Nagy. Benchmarking and standardization of intelligent robotic systems. In *14th International Conference on Advanced Robotics (ICAR 2009)*, Munich, Germany, June 2009.

[11] NIST. NIST Response Robot Evaluation Exercise. Search and Rescue: Texas Engineering Extension Service (TEEX), November 2008.

[12] K. Pathak, N. Vaskevicius, J. Poppinga, M. Pfingsthorn, S. Schwertfeger, and A. Birk. Fast 3d mapping by matching planes extracted from range sensor point-clouds. In *International Conference on Intelligent Robots and Systems (IROS09)*, St Louis, MO, USA, September 2009. IEEE.

[13] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L.-E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. van Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney. Stanley: The robot that won the darpa grand challenge: Research articles. *J. Robot. Syst.*, 23(9):661–692, 2006.