User-Specified Performance Metrics for Autonomous Robots in Warehouse Logistics

Lutz Frommberger, Torsten Hildebrandt, Bernd Scholz-Reiter

***Abstract*— Autonomous robots rely on a large variety of different skills that all contribute to solving a concrete problem. For any of these skills, performance measures exist. It is not always clear, however, how expressive these measures are with regard to the overall performance of the robotic system in a particular task. In this paper, we discuss performance metrics for autonomous robots on an example from warehouse logistics. We point out that individual metrics for lower-level subtasks might have limited value for assessing the whole system from a logistics perspective, also depending on the task to be solved. Thus, we argue for formalized user-defined performance measures on a high level of abstraction that allow a human operator to assess the things that really matter for the problem at hand. We exemplify this by showing results from an ongoing case-study with a surveillance robot in a simplified warehouse setting.**

I. INTRODUCTION

For a logistics company running a warehouse, the use of robots is an economic decision: They are employed if it either pays off monetarily or yields other benefits. In a warehouse context, potential savings can be due to reduced storage times in the warehouse, shorter traverse paths, higher throughput of goods, etc.. Looking at these parameters, it can easily be evaluated if the use of a robotic system pays off in contrast to not using it. This comparison, however, can only be performed after the robotic system has been implemented—but a company would want to answer the question of cost-effectiveness *before* paying for implementing a new system.

In this paper, we particularly look at surveillance robots. To complete their mission, such robots have to perform several different tasks, such as mapping and localization of the environment, navigation within the warehouse, cooperation of robot teams, identification of goods, detection of storage processes, etc.. For each of these tasks, performance measures can be defined, and strategies can be adapted to optimize these measures. The crucial point here is that the named tasks are not independent. On the one hand, better performance in one task will increase the performance of another task (for example, better object recognition might have a positive effect on mapping). On the other hand, the opposite can be the case, and an improvement in one task can lead to a loss of performance in another; exploration vs. exploitation is a classical example here. Thus, optimizing the overall performance under cost constraints results in a multidimensional optimization problem where the impact of a single system component is hardly tractable.

We discuss the problem of evaluating the performance of a surveillance robot system in a concrete scenario in the field of warehouse logistics. The questions posed are of a general kind, however, and apply to many more problems where multiple (or single) robots are employed in a logistics context.

In the following, we first describe the warehouse scenario we use as an ongoing example in this paper. In Section III, we investigate performance measures both from a logistic and a robotic point of view and exemplify shortcomings of common benchmarks. Then, we argue for user defined abstract performance descriptions in Section IV in order to focus on the relevant information for the task at hand. To exemplify this, Section V reports on an evaluation of process detection within a simplified warehouse scenario before we end with a summary.

II. SCENARIO DESCRIPTION

In this paper, we investigate the problem of measuring the performance of a robotic system by examining a concrete project dealing with the use of surveillance robots (see Figure 1) in warehouse logistics in order to enable semi-automatic logistic optimization. We address the problem of understanding so-called *chaotic* or *random-storage warehouses*, characterized by a lacking predefined spatial structure, that is, there is no fixed assignment of storage locations to specific goods. Thus, storage processes are solely in the responsibility of the warehouse operators and basically not predictable—goods of the same type may be distributed over various locations. This makes it a hard problem for people aiming at *understanding* the in-warehouse processes, e.g., different storage patterns. Knowledge about these storage patterns is, however, of critical importance to understand what is going on in the warehouse, and this is a prerequisite to enable an operator to optimize the warehouse.

In this scenario, we have a solid set of background knowledge, for example, we know that storage processes always follow the same schemata involving *functional zones* within the warehouse. But the concrete details—for example, the concrete locations of these zones—are unknown and can even change over time. Using autonomous robots as a minimally invasive means to observe in-warehouse processes in order to support a logistic optimization of a chaotic

Lutz Frommberger is with the Cognitive Systems Group, Dept. of Mathematics and Informatics, University of Bremen, Germany, email: lutz@informatik.uni-bremen.de

Torsten Hildebrandt and Bernd Scholz-Reiter are with the Dept. of Production Planning and Control, University of Bremen, Germany, email: hil@biba.uni-bremen.de, bsr@biba.uni-bremen.de

Work carried out in this paper was supported by the APF focus project program of the University of Bremen, Grant Nr. 011 600, and the collaborate research centers SFB/TR 8 “Spatial Cognition” and SFB 637 “Autonomous Logistic Systems”, funded by the German Research Foundation (DFG).

Fig. 1. A surveillance robot in a warehouse scenario setup

warehouse was proposed in [14]. The information provided by the robot shall allow for posing queries about observed spatio-temporal activities (such as "How often have goods been relocated within the storage zone?") as well as about regions in space (e.g., "Which areas in the warehouse have been used as a buffer zone?"). Answering such queries is an important step towards logistic optimization.

The goal of this project is to specify and detect high-level logistic processes inside the warehouse and to make them available to build a logistic simulation that can then be used for optimization purposes. This optimization is not within the scope of this paper; we restrict ourselves to the process detection based on observations from robots.

III. PERFORMANCE MEASURES

Both logistic systems as well as the robot systems are complex systems and their design involves the proper choice of various design options. Expressed more formally it requires the choice of specific design options for a logistic (robotic) system out of a certain "design space" $d_L \in D_L$ ($d_R \in D_R$). Examples of such design choices for the robotic system can be which mapping algorithms to use, the object recognition to employ, or how many robot units to use. Design choices for a warehouse system are, e.g., the order picking policy used, or the layout of storage areas and whether to use fixed assignments of certain kinds of goods to specific storage locations.

These design options determine the performance characteristics (including cost) of the system. Using $p_L \in P_L$ ($p_R \in P_R$) to denote the performance vectors of the logistic (robotic) system, we can define a function to derive the mapping from the design space to the "performance space" $f_L : D_L \rightarrow P_L$ ($f_R : D_R \rightarrow P_R$). To actually determine this mapping and find the potentially complex relationships between design options and performance will typically require simulation, or even physical experimentation for both types of systems.

The semi-automatic, robot-supported optimization of warehouse systems intended in this project can now be seen on an abstract level as a function $g_{\text{opt}} : P_R \rightarrow D_L$. The use of the robotic system with performance characteristics $p_R \in P_R$ allows the surveillance of the warehouse system and enables a logistics expert to simulate the system, and find new, potentially better design choices $d_L \in D_L$ for the logistic system.

Therefore we get logistic performances before (d_L) and after (d'_L) the robot-supported optimization. Whether d'_L is now better, and therefore the use of the robotic system and subsequent optimization is successful, depends on the preferences of the company running the warehouse, as the problem is a multi-objective decision problem. Let $\underset{u}{\succ}: P_L \times P_L \rightarrow \{0, 1\}$ denote the preference (or utility) function of this company, which is 1 iff the first performance vector is preferred to the second, or 0 otherwise. Then the use of the robot-supported optimization was successful, if $f_L(d'_L) \underset{u}{\succ} f_L(d_L)$, with $d'_L = g_{\text{opt}}(f_R(d_R))$.

A. *A logistic view on performance*

In general measuring the performance of a logistic system is difficult [5]. Depending on the level of abstraction, time horizon, as well as potential recipients, relevant information differs. Usually an internal and external view on logistic performance can be defined [19], reflecting different conflicting goals of running a logistic system. The external, customer-oriented view focuses on logistic quality, i.e., short delivery times and high adherence to delivery dates, whereas the internal view tries to quantify logistics efficiency trying to ensure cost-efficient operation (examples include capacity utilization, stock levels, turnover frequency). Thus, logistic performance is inherently multi-dimensional. There are, however, attempts to aggregate various performance indicators in hierarchic performance measurement systems [25].

Logistic systems are complex, highly dynamic systems, so insights into the operation and assessing the impact of changes to such systems are often not feasible using analytic approaches, thus requiring simulation models. Using such models allows to optimize certain aspects of the system, such as the order picking strategy in a warehouse [6], either manually by domain experts or automatically using simulation-based optimization [11].

Assessing the success of such optimization is usually evaluated focusing on some low level performance measure and its improvement. Example measures for optimizing warehouse operations are throughput of picking orders, or the average/maximum time till fulfillment of picking orders. This, however, is only a simplification, and performance measures can rarely be treated in isolation. Therefore, to decide which of several options to prefer depends on a user's preferences (denoted with $\underset{u}{\succ}$ in the previous section) in order to compare otherwise unrelated performance measures (such as costs, and, e.g., order fulfillment time) with each other. Without knowing this preference function only clearly inferior options can be ruled out, i.e., those worse in every dimension, which are therefore Pareto-dominated.

In our scenario, these logistic measures only refer very indirectly to the performance of the observing robot, as it is only providing information for the optimization process and, thus, is not directly involved into the operation of the warehouse. It only provides information that is then used for warehouse optimization later on. To judge the quality of the observations, we need to consider different benchmarks.

B. A robotics view on performance

For the given warehouse scenario, the robots have to follow a complex behavior that tackles many different robotics tasks: The environment needs to be mapped and the robots have to localize themselves in the generated map; the environment needs to be explored continuously because of its dynamics; objects have to be tracked and to be identified; activities happening to those objects have to be recognized and categorized. For all these subtasks, reasonable performance measures exist. In the following, we provide an overview by giving a (non-exhaustive) set of examples.

1) Localization and Mapping Quality: The general approach to evaluating the success of mapping and localization is to calculate an absolute error over the distance between prediction and actual position of the robot or features in the map (the *ground truth*). For examples, see [20] or [7]. The main problem arising here is that the ground truth might not be easily available. For situations where the precise robot position cannot be retrieved easily for comparison, more semantically driven approaches have been proposed (e.g. [9]). Recent work focuses more on challenges emerging from robot use in the field. For example, [28] proposed suitable benchmarks for 6-D outdoor SLAM. The measures mentioned up to now are designed for grid based maps; for topological maps, an approach is to define an error measure of how well a generated road graph fits into the geometric map, e.g. [26].

2) Multi-robot coordination and exploration: The performance metric for coordinated multi-robot exploration usually is the time in which they manage to finish a coverage task, that is, to map the whole environment (e.g., [4], [24]). Coordination mechanisms are evaluated implicitly with regard to coverage performance. Often, evaluation stops at the point of complete coverage. When it comes to repeated coverage, as in sweeping tasks or our warehouse surveillance task, evaluation usually becomes anecdotal [2] or performance is proven by analytic assurances [10].

3) Object recognition: Object recognition[1] is a field that is intensively benchmarked. To rate the detection success, methods are usually evaluated over standardized object libraries such as the "Columbia object image library" (COIL) [18] that has been used in various seminal approaches (e.g., [23], [3]), or the Caltech 101 or Caltech 256 data sets [13]. While this ensures very comparable results, it has been pointed out that results relying on these benchmarks can be highly misleading and might lead into very wrong research directions [21]. For many specific tasks, more specific benchmarks are developed, for example for pedestrian detection [8].

4) Activity recognition: Activity recognition or process recognition tasks are usually evaluated counting the detection rate against a human-annotated ground truth. While there had been no standard metric to compare such systems 5 years ago [17] and metrics borrowed from related fields often failed to provide meaningful insights [27], recent developments have led to quite some progress with regard to performance metrics [27]. Especially in the field of human activity recognition, standard benchmarks have been set up [12], and even an "Activity Recognition Challenge" is running this year [1].

C. Shortcomings

We have seen that a multitude of metrics exist for evaluating the performance of a robotic system with regard to specific subtasks. In a robotics application in logistics, however, several of these subtasks contribute to the overall performance. One might be tempted to try to increase the performance of the robotic system by optimizing the performance in any subtask. Of course, this does not work as hoped for, as the subtasks are not independent. In the following, we give two examples.

1) Mapping Quality vs. Exploration: Mapping of the environment is an important component of most robotics applications, as it is needed for navigation and localization of the machine. Assessing the quality of a generated grid map is not a hard problem (see Section III-B.1). But apart from the sensors and algorithms used, the mapping quality is also affected by the navigation speed of the mapping robot, higher speed might easily lead to poorer results. But the question is up to which point a high-quality grid map is needed for the task at hand. Topological approaches with local navigation controllers that rely on less exact data have shown great success and robustness as well [16, e.g.].

Faster navigation and, following from that, a slight decrease in mapping quality will pay off in faster exploration of the environment and this creates more opportunities to detect activities. In surveillance tasks, this might be more important than a detailed and exact grid map, so maximizing the performance of the mapping system might be the wrong idea at some point.

2) Activity Frequency vs. Activity Diversity: When surveying the warehouse, we would like to detect at the best *all* activities involving the movement of goods that happen in there—which is not realistic, unless we place a stationary robot everywhere. The robots have to move within the warehouse to find activities and, thus, they will miss some. Let us assume our measure would be to count the percentage

[1] Object recognition is not a serious issue in the given warehouse scenario, as we rely on uniquely identifiable features such as RFID tags.

of detected activities over time. Then, the best strategy according to this measure would be that a robot would remain at a location where many activities are going on (perhaps the entrance, where all goods will be located when entering the warehouse). This extreme case would create a very clear picture of this one location (a *hotspot*), while other locations remain unexplored. In this case, we will miss important information. For example, it will remain unclear whether some storage location is not used at all or rarely used—and under which circumstances. This distinction might be a relevant detail for the optimization purpose.

IV. Retrieving Essential Information

A. Describing processes on different levels of abstraction

The difficulty when assessing different performance measures in the overall scenario is how relevant the performance metrics are for the overall process. On the lower sensory level, this question is hard to answer. As argued for in Section III-C, an improvement with regard to one measure might even lead to a performance drop for the overall system. The important question is: How relevant is the measured behavior for the targeted outcome?

To be able to utilize the acquired sensory input coming from the robot in a reasonable way, the sensory data is subject to interpretation, leading to different levels of abstraction for the perceived data. Measured distances are transformed into a map, dynamic changes within the map are interpreted as activities, certain patterns of activities have a specified meaning as an abstract concept—for example, they can be interpreted as a redistribution of the good within the warehouse. Such higher-level concepts allows an expert in the logistics domain a much clearer assessment of how relevant the measured parameters are. While, for example, it is doubtful what an increased range of sensor perception will contribute to the overall success of the system, the percentage of correctly detected redistributions in a warehouse can much easier be related to its importance for the overall performance. Clear abstract descriptions of high-level processes especially become important when human experts are included in the assessment process.

In the framework given in Section III, changing one or more options in the flat vector of design options (d_R) in the hope to improve the overall logistic performance P_L usually results in cumbersome trial-and-error experiments with a following long simulation phase. Instead of doing so we argue for having (limited) set of design options that are integrated to contribute to higher level concepts that become part of the performance space P_R.

B. User-defined Queries for Individual Performance Assessment

Depending on the task at hand, *different* processes become relevant. While it might once be the redistributions we are interested in, it might also be the information about whether different goods are moved out of the warehouse together. Thus, it is reasonable that an operator can specify relevant in-warehouse processes according to his needs. For

Fig. 2. The idealized warehouse scenario used for evaluation of the symbolic process detection approach

the described warehouse surveillance task, Kreutzmann et al. show how this can be done by using a formal language [15], namely linear temporal logic (LTL) [22]. For example, without going into details in LTL, a *redistribution* operation was specified like this:

$$\text{Redistribution}_{G,L_i,L_j} = \text{at}(G, L_i) \wedge \text{in}(L_i, S) \rightarrow \Diamond\big(\text{at}(G, L_j) \wedge \text{in}(L_j, S)\big) \quad (1)$$

This defines that we have a redistribution when a good G is observed at a location L_i (at(G, L_i)) and this location lies within the storage zone S (in(L_i, S)) and at some point in time later (specified by the modal operator $\Diamond$) G is observed at another location L_j (at(G, L_j)) which also lies within the storage zone (in(L_j, S)).

More abstract processes specified like this can then also be used for evaluation by judging the success with regard to exactly this process with a metric specified by a human operator, that is, it becomes part of the performance space P_R of the robotic system. Various different specifications of processes and measures can be specified and evaluated concurrently (e.g., having both measures for exploration and exploitation) such that it is, similar to the utility function $\underset{u}{\succ}$, the system operator's preference which measures to optimize in order to achieve a better overall performance.

In the following evaluation, we will introduce the concept of *histories* of goods, a concept that is not really relevant for the understanding of the warehouse but has been specified as a measure to integrate different facets of the robot's performance.

V. An Example Evaluation within the Warehouse Scenario

In this section we report on a first experimental evaluation performed for the LTL-based process recognition approach, as described in [15]. The goal was to measure the success of the proposed process detection mechanism in an idealized warehouse scenario set up in the lab (see Figure 2).

In the experiment, cradles were moved around in the warehouse between pre-defined zones while the robot was

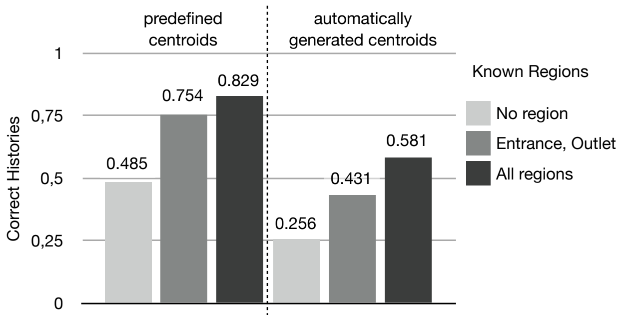

Fig. 3. Evaluation results showing the rate of correct history detections. Diagram taken from [15].

steered within this scene and mapped the position of the cradles which it could identify by unique optical tags attached to them. The processes to be detected had been specified in LTL, as described in the previous section. Observations made by the robot's sensory system were abstracted and translated to symbolic expressions (like at(...) or in(...)), and a process was detected if we can find a model (based on these symbolic observations) that satisfies the corresponding formal specification.

The question was now what to measure when evaluating the approach. An obvious metric would be to count the percentage of correctly detected in-warehouse processes. However, this would veil the information which processes had been detected well and which not. A fairly high percentage of correctly detected observation could eventually camouflage the fact that one particular process had hardly been correctly observed at all. So the evaluation considered correctly observed *histories* of goods. The *history* of a good is the sequence of different processes that involves one particular good. Looking at histories ensures that the flow of goods over the whole warehouse is monitored and assessed. A correctly determined history ensures that *all kinds of processes* can be recognized and the whole spatial range of the warehouse is covered. Histories can also be formulated as LTL formulas. A history is correctly identified if temporal order and number of processes match the ground truth.

Figure 3 shows results from the experiment under different amounts of background knowledge. In particular, the success was evaluated with, without, and with only partial knowledge of the functional regions. Also, the performance of the clustering algorithm that translates measured positions of goods automatically to a smaller, discrete set of qualitative abstract locations was compared to a clustering algorithm where the cluster centers had been (reasonably) pre-defined. The results reveal the major problem of the implementation so far: with automatic clustering, detection rates of histories significantly dropped. Especially, this problem became evident for the case of only knowing two of the functional regions (entrance and outlet), which is a most realistic assumption in this scenario. With pre-defined clustering, the proposed algorithm could infer the knowledge about regions very well (detection rate only dropped from 82.9% to 75.4%), but with automatic clustering it dropped from 58.1% to 43.1%. The used clustering algorithm was known to be quite straight-forward, but it was assumed that its quality was sufficient. The given experiment disproved this assumption by evaluating its impact on high-level process detection. Thus, the evaluation approved one of the main goals of the approach: to show that complete knowledge about the whereabouts of the functional regions is not really necessary and can be deduced by formal reasoning—but only with a sufficiently high quality of the clustering method. Improving this quality is a reasonable next step to improve the overall performance of the system.

VI. SUMMARY

In this paper, we have investigated how to evaluate the performance of a surveillance robot system in order to gather information for logistic optimization a warehouse scenario. Several options exist to improve the robot's behavior in certain subtasks, and all of the subtasks have individual performance measures. The impact of different design options to overall logistic performance measures is hard to understand, as they mutually affect each other. We have exemplified shortcomings of current robotics metrics with regard to their contribution to a complex system and argued for an integrating approach where users can specify performance metrics on a higher level of abstraction such that they can be related to the overall performance more easily. Finally, we have exemplified this by showing evaluation results from an ongoing case study that was able to identify shortcomings in one particular part of the system by evaluating higher-level concepts.

REFERENCES

[1] Activity recognition challenge. `http://www.opportunity-project.eu/challenge`.
[2] M. Ahmadi and P. Stone. A multi-robot system for continuous area sweeping tasks. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1724–1729, 2006.
[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:509–522, 2002.
[4] W. Burgard, M. Moors, D. Fox, R. Simmons, and S. Thrun. Collaborative multi-robot exploration. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 476–481, 2000.
[5] Garland Chow, Trevor D. Heaver, and Lennart E. Henriksson. Logistics performance: Definition and measurement. *International Journal of Physical Distribution & Logistics Management*, 24(1):17–28, 1994.
[6] R. Dekoster, T. Leduc, and K. Roodbergen. Design and control of warehouse order picking: A literature review. *European Journal of Operational Research*, 182(2):481–501, 2007.
[7] F. Dellaert, D. Fox, W. Burgard, and S. Thrun. Monte carlo localization for mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1322–1328, 1999.
[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 304–311, 2009.
[9] T. Duckett and U. Nehmzow. Mobile robot self-localisation and measurement of performance in middle-scale environments. *Robotics and Autonomous Systems*, 24(1-2):57–69, 1998.
[10] Yehuda Elmaliach, Noa Agmon, and Gal Kaminka. Multi-robot area patrol under frequency constraints. *Annals of Mathematics and Artificial Intelligence*, 57:293–320, 2009.
[11] M. Fu. Simulation optimization. In *Proceedings of the 2001 Winter Simulation Conference*, pages 53–61. IEEE Press, 2001.

[12] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *In ICCV*, 2005.

[13] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007.

[14] Torsten Hildebrandt, Lutz Frommberger, Diedrich Wolter, Christian Zabel, Christian Freksa, and Bernd Scholz-Reiter. Towards optimization of manufacturing systems using autonomous robotic observers. In *Proceedings of the 7th CIRP International Conference on Intelligent Computation in Manufacturing Engineering (ICME)*, June 2010.

[15] Arne Kreutzmann, Immo Colonius, Lutz Frommberger, Frank Dylla, Christian Freksa, and Diedrich Wolter. On process recognition by logical inference. In *Proceedings of the 5th European Conference on Mobile Robots*, September 2011.

[16] Benjamin Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119:191–233, 2000.

[17] D. Minnen, T. Westeyn, T. Starner, J. Ward, and P. Lukowicz. Performance metrics and evaluation issues for continuous activity recognition. *Performance Metrics for Intelligent Systems*, 2006.

[18] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (COIL-20). Technical report, Department of Computer Science, Columbia University, New York, 1996.

[19] Peter Nyhuis and Hans-Peter Wiendahl. *Fundamentals of Production Logistics: Theory, Tools and Applications*. Springer, 2008.

[20] S. Oore, G. E. Hinton, and G. Dudek. A mobile robot that learns its place. *Neural Computation*, 9(3):683–699, 1997.

[21] Nicolas Pinto, David D. Cox, and James J. DiCarlo. Why is real-world visual object recognition hard? *Computational Biology*, 4(1), 2008.

[22] Amir Pnueli. The temporal logic of programs. In *Proceedings of the 18th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 46–57, 1977.

[23] M. Pontil and A. Verri. Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–646, 1998.

[24] R. Simmons, D. Apfelbaum, W. Burgard, D. Fox, M. Moors, S. Thrun, and H. Younes. Coordination for multi-robot exploration and mapping. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 852–858, 2000.

[25] Rien Ploos van Amstel and Guido D'hert. Performance indicators in distribution. *The International Journal of Logistics Management*, 7(1):73–82, 1996.

[26] Jan Oliver Wallgrün. Qualitative spatial reasoning for topological map learning. *Spatial Cognition and Computation: An Interdisciplinary Journal*, 10(4):207–246, 2010.

[27] J.A. Ward, P. Lukowicz, and H.W. Gellersen. Performance metrics for activity recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):6, 2011.

[28] O. Wulf, A. Nüchter, J. Hertzberg, and B. Wagner. Benchmarking urban six-degree-of-freedom simultaneous localization and mapping. *Journal of Field Robotics*, 25(3):148–163, 2008.