Physical Interactive Localization Learning

Muhammad Haris, Mathias Franzius, Ute Bauer-Wersing

2022

Preprint:

This is an accepted article published in 2022 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO). The final authenticated version is available online at: https://doi.org/[DOI not available]

Physical Interactive Localization Learning

Muhammad Haris[1] and Mathias Franzius[2] and Ute Bauer-Wersing[1]

***Abstract*— Localization is fundamental for mobile robots, especially in unconstrained outdoor environments. Earlier work showed unsupervised localization learning on landmarks to be suitable for large-scale scenes. However, this relied on hand-labeled data to train a CNN for recognizing landmarks. We propose a new approach that allows a robot to learn landmarks for localization with a human cooperatively. This approach uses pre-trained detectors of common objects for learning new landmarks in a scene, requiring only minimal human supervision. Hence, the method bootstraps the landmark learning process and removes the need to manually label large amounts of data. We present localization results using the learned landmarks in simulated and real-world outdoor environments and compare the results to models based on complete images and PoseNet. The landmark-based localization shows improved performance than the baseline methods in challenging scenarios. Our results further show that localization accuracy increases with the number of learned landmarks. The human teacher has complete control over selecting new landmarks, which allows learning unique, robustly detectable, and semantic landmarks.**

I. INTRODUCTION

Localization determines the position of a robot in an environment (i.e., *where am I?*) w.r.t a known map. In contrast to GPS or laser scanners, cameras are a widespread choice for localization because of their small size, low power consumption, and low cost. The field covers various algorithms developed over the last two decades that employ image features [1], [2], train neural networks [3], [4], or derive inspiration from neurobiological systems [5], [6], [7]. A set of methods combine object detection with localization [8], [9], [10], [11] to generate maps that contain scene semantics for achieving high-level intelligent robot behavior besides precise localization.

In addition to other autonomous mobile agents, the ability to localize in an environment is also a fundamental requirement for social robots. A social robot can be defined as a physically embodied autonomous agent engaging in social interactions with humans by communicating, cooperating, and making decisions [12], [13]. The two primary applications for household robots today are cleaning and lawn mowing, which are service robots assisting humans in dirty, dull, or repetitive tasks. Here, the service robot is intended to provide services to humans and not vice versa. However, practically, due to limited robot capabilities, humans still have to provide services to their robots, e.g., by installing infrastructure like border wires or saving robots that are stuck or out of battery. Many of the missing capabilities are due to the limited localization abilities of the robot. The resulting interactions are typically unintuitive and unpleasant but unavoidable.

Here, we propose a novel approach to making these interactions with service robots more intuitive, cooperative, and enjoyable. In our application example, a lawnmower robot comes with factory-provided recognition for a limited set of landmarks that can be used for localization. However, more detected unique landmarks in an environment improve localization and thus improve work efficiency and the probability of successful homing and recharging, decreasing the likelihood of leaving a defined work area or colliding with static objects. Providing sufficient landmark detectors a priori for gardens is challenging since suitable landmarks should be long-term detectable and unique. In a garden, this could mean, for example, distinguishing a specific tree from all other trees in the environment. Thus, the landmarks need to be learned interactively on site. The most prominent state-of-the-art method to learn object or landmark detectors is manual labeling, e.g., having humans draw rectangles around such objects in thousands of images. This method is very time-consuming and costly.

Our proposed method replaces this task with an intuitive human-robot interaction scheme. We propose a method to bootstrap the label generation process using physical object instances with pre-trained visual detectors (e.g., MS-COCO objects [14]) as a labeling tool to learn new landmarks for localization. We will refer to objects with pre-trained detectors as *anchors* while the objects or regions to-be-learned as *landmarks*. A human interactor places a physical object as an anchor next to a region to teach the robot landmark recognition. The robot now collects views from varying perspectives, detects the anchor, infers the landmark position from the anchor position, and generates the annotated image data for learning the new landmark (fig. 1a). This bootstrapping process allows efficient and fast generation of labeled training data compared to a cumbersome hand-labeling process. Moreover, bootstrapping from anchors enable us to generate training data for many landmarks in the scene and adapt localization accuracy according to the needs. We perform experiments in simulated and real-world outdoor environments and report localization performance on anchors and landmarks. We use PoseNet [4] and SFA localization on complete images [15] as baseline methods. PoseNet is an end-to-end learning-based visual localization method that directly regresses an image's pose. Like SFA, PoseNet also has an offline learning phase (mapping) that allows a straightforward comparison.

[1]Muhammad Haris and Ute Bauer-Wersing are with the Faculty of Computer Science and Engineering, Frankfurt University of Applied Sciences, 60318 Frankfurt, Germany muhammad.haris@fb2.fra-uas.de

[2]Mathias Franzius is with the Honda Research Institute Europe GmbH, 63073 Offenbach, Germany

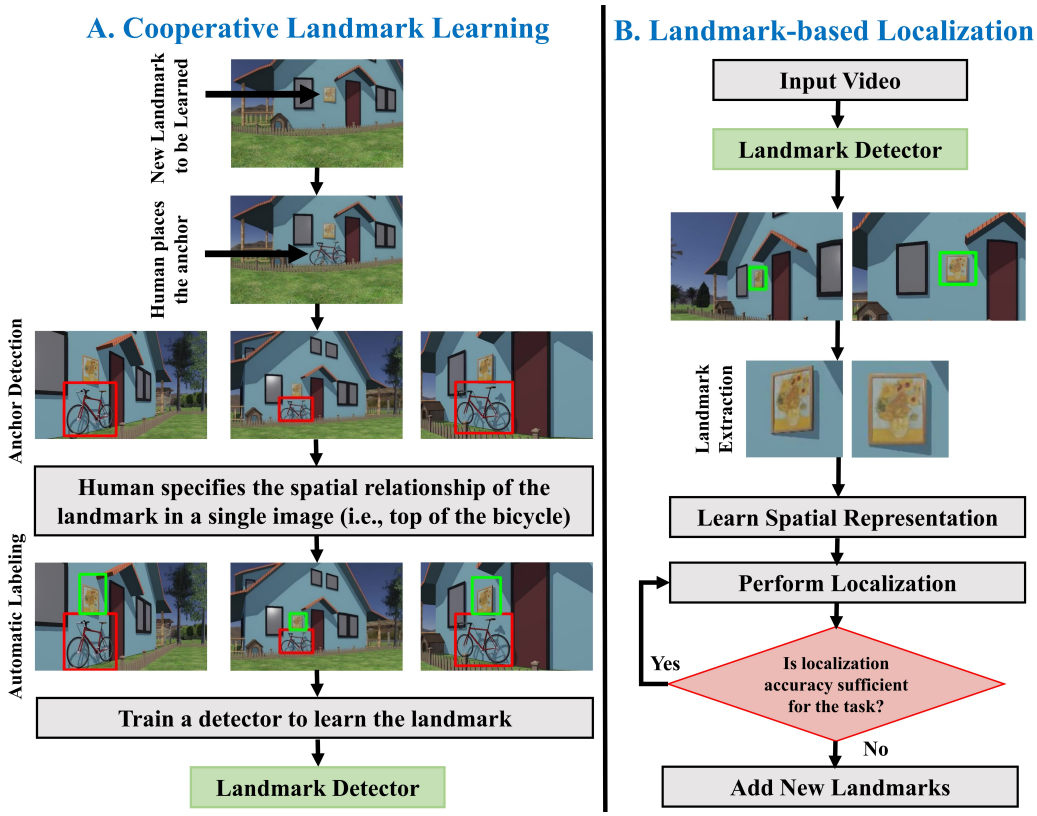

Fig. 1. **Cooperative Landmark Learning and Localization:** (a) A human places a pre-trained anchor (i.e., a bicycle) in a scene to facilitate labeled data generation for learning a new landmark (i.e., painting). The robot then derives and collects imagery from different viewpoints. Afterwards, the system applies the pre-trained detector to recognize the bicycle in the images. A human now specifies the spatial relationship of the new landmark (i.e., top of the bicycle) in a single image. Next, the system generates the labeled training data from all available images where the bicycle is detected. Finally, the system uses the generated labeled data to train a CNN for landmark detection. (b) The robot uses the learned detector to identify the new landmarks in the images collected for mapping and localization phases. The next step extracts the landmarks from the images and resizes them to a fixed size of 120 x 120 pixels. The robot then learns spatial representation relative to the landmark and performs localization. If the localization accuracy is not sufficient, the robot can improve the performance by learning multiple landmarks w.r.t to an anchor or can ask a human to place new anchors (a) in the regions where a higher performance is desired.

II. METHODS

A. Cooperative Landmark Learning

The core idea in this work is to use physical objects in a scene with pre-trained visual detectors as a labeling tool for learning new landmarks. It can be achieved by using or placing such known objects in spatial relation (e.g., next to, or under) landmark instances that shall be learned to be detected. Fig. 1a shows the steps to learn new landmarks for localization cooperatively with a human. Consider our goal is to learn the painting on the wall (fig. 1a) as a new landmark. A human interactor places the anchor (i.e., a bicycle) below it to teach the robot landmark recognition. The robot now traverses in the environment to collect views from varying perspectives and applies a pre-trained bicycle detector (i.e., YOLOv3 [14]) to detect its instance in the recorded images. A human now only needs to indicate the spatial relation of the bicycle relative to the landmark once to generate labeled training data for this landmark from all recorded images. Thus, the method bootstraps the landmark learning process and removes the need to label large amounts of data manually. In this work, we use a 2D offset (i.e., above, below) to derive a landmark relative to an anchor. Please note that it is possible to derive multiple landmarks from a single anchor, making it straightforward to scale the system without increasing the number of anchors. The system then uses the annotated image data for the new landmark and trains a CNN to learn it. The anchor is no longer required after the landmark learning phase and thus can be removed optionally from the scene. The robot uses the trained landmark detector for the localization and mapping phases. This process can be repeated to learn more landmarks. A human has complete control over the selection of new landmarks. Thus, landmarks can be made unique, robustly detectable, and semantically meaningful, allowing task-relevant localization accuracy in different regions and interactive learning in a playful way.

B. Landmark-based Localization

1) Slow Feature Analysis: To learn the robot's position in 2D space, we use Slow Feature Analysis (SFA) as introduced in [16]. It transforms a multidimensional time series $\mathbf{x}(t)$, in our case images, along a trajectory, to slowly varying output signals. The objective is to find instantaneous scalar input-

output functions $g_j(\mathbf{x})$ such that the output signals

$$s_j(t) := g_j(\mathbf{x}(t))$$

minimize

$$\Delta(s_j) := \langle \dot{s}_j^2 \rangle_t$$

under the constraints

$$\begin{aligned} \langle s_j \rangle_t &= 0 \text{ (zero mean)}, \\ \langle s_j^2 \rangle_t &= 1 \text{ (unit variance)}, \\ \forall i < j : \langle s_i s_j \rangle_t &= 0 \text{ (decorrelation and order)} \end{aligned}$$

with $\langle \cdot \rangle_t$ and $\dot{s}$ indicating temporal averaging and the derivative of s, respectively. The Δ-value is a measure of the temporal slowness of the signal $s_j(t)$. It is given by the mean square of the signal's temporal derivative, so small Δ-values indicate slowly varying signals. The constraints avoid the trivial constant solution and ensure that different functions g_j code for different aspects of the input. We use the MDP [17] implementation of SFA, which is based on solving a generalized eigenvalue problem.

2) Localization Learning on Landmark Views: Fig. 1b shows the steps to acquire landmark views and, consequently, use them to learn spatial representation. The input is an image sequence obtained by the robot exploration phase in an environment. The system then applies the learned detector from the previous step to recognize the landmark in the collected images. Afterwards, the system extracts the landmark views from the images and rescales them to a fixed size of 120×120 pixels. The output of this step generates an image stream that contains landmark views. The procedure is the same for the images of training (mapping) and test (localization) phases.

The next step uses landmark views to extract the spatial representation of an environment. We employ a hierarchical, feed-forward, 4-layer SFA network [11] that learns the position encoding relative to a landmark in an unsupervised way. After training, the next step uses the learned representation and the robot's odometry information (x, y) to learn a regression function that maps the network's output to metric space. This step is done solely for evaluation purposes. In the case of multiple landmarks, we train a separate SFA network for each landmark. Therefore, each landmark will be an independent estimator of the robot position (x, y) after the mapping phase.

The robot uses the learned position estimator to estimate its location (x, y) in the localization phase. If the robot detects that the localization performance is insufficient to execute a task successfully, it can ask a human to put a known object at some position so that it can learn new landmarks. Hence, the proposed method allows to interactively learn new landmarks in a scene with a human in the loop and scale localization accuracy according to the needs.

III. EXPERIMENTS

This section presents localization results from simulated and real-world outdoor environments. We derive a single landmark for each anchor present in a scene by setting a fixed 2D offset and then train a CNN to learn the landmarks. If the anchor and the landmark are not within the same plane, a simple 2D offset will not capture a semantic region but a subset of the scene's viewing space. However, our experiments suggest that these views can be classified with a CNN, and we expect no degradation of localization accuracy on such views. After the learning phase, the system uses the learned landmarks for localization. We also perform localization w.r.t anchors, plain SFA [15], and PoseNet [4] for comparison.

A. Simulated Experiments

We performed the experiments in a simulated garden with an area of 18×18 meters. A robot randomly traverses the area to record images for the training set and then samples a regular grid to collect the test set. During traversal, it captures panoramic views of size 3600×600 pixels. The training and test trajectory consist of 15000 and 1250 images, respectively. The virtual environment consists of three different CNN-detectable anchors (i.e., a bicycle, a car, and an umbrella). The system then learns one landmark relative to each anchor by training a CNN. After training, the learned landmarks achieve similar recognition rates as pre-trained anchors (c.f. table I). The system then uses the landmarks to perform localization. Table I reports the median localization performance w.r.t anchors, landmarks, and the baseline methods. The median metric is used for comparison to remove the effect of outlier position estimates (x, y) obtained with the baseline method (i.e., PoseNet). The combination of landmarks achieves superior localization performance than the baseline methods in this experiment.

B. Real-world Experiments

We performed the real-world experiments in two garden-like outdoor scenes that differ w.r.t to their size and complexity. The small-scale garden has an area of 9×13 meters, while the large-scale garden has an area of 31×20 meters. We used a lawnmower robot (fig. 2a) equipped with a fisheye lens that captures omnidirectional views of size 2880×2880 pixels. During recordings, it traverses along a fixed boundary wire (fig. 2c, 2d) to record the imagery for the experiments. We generated the robot's ground truth position (x, y) with a structure-from-motion (SfM) software[1] and used it to evaluate the localization performance. We recorded three different image datasets from each environment, where we used one to train the system and the other two as test sets. These datasets vary w.r.t to daytime, lighting conditions, and dynamic scene changes. Figure 2b shows an example of the anchor (i.e., a suitcase) and the learned landmark (i.e., an electrical cabinet) present in one of the outdoor environments.

[1] https://www.agisoft.com/

TABLE I

Localization Results on Simulated Data: Median localization performance w.r.t anchors, learned landmarks, and the baseline methods. The table further reports detection rates of anchors and learned landmarks. The combined detection rate of 100% indicates that at least a single landmark is present at any given test location. Localization w.r.t learned landmarks outperforms the baseline methods in this experiment.

	Landmark-based Localization					**Image-SFA**	**PoseNet**
	Id_1	**Id_2**	**Id_3**	**Combined**	**%**		
Anchors	0.47m [100 %]	0.43m [99 %]	0.16m [99 %]	0.22m	100	0.26m	0.21m
Learned	0.18m [99 %]	0.27m [99 %]	0.27m [97 %]	**0.16m**	100		

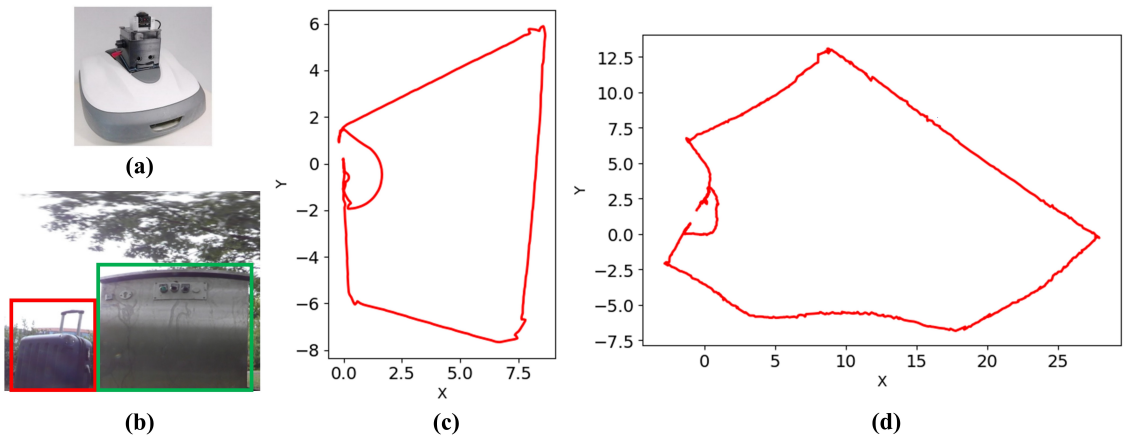

Fig. 2. **Experimental Setup:** (a) An autonomous lawn mower robot with a fisheye camera was used for the experiments. (b) shows an example of a pre-trained anchor (i.e., a suitcase) and a derived landmark (i.e., an electrical cabinet) relative to the anchor in a real-world environment. (c) and (d) show the robot's traversed trajectories in one of the recording sessions from both gardens, respectively.

1) Small-scale Garden: The training and test sets include images from a similar robot trajectory (fig. 2c) but different environmental conditions (i.e., daytime, weather, and dynamic objects). The training set consists of 1138 images, while the two test sets have 1091 and 1109 images. We used 25% of the data (subsampled from the training set) as the validation set to evaluate PoseNet performance during training. After the training phase, PoseNet achieves a localization accuracy of 0.06m on the validation set. The small garden has three pre-trained anchors, which the system uses to learn new landmarks for localization. Table II shows the detection rates and the median localization performance w.r.t to the learned landmarks and the baseline methods. The use of a single landmark enables coarse localization in an environment. However, adding more landmarks leads to a higher performance similar or better than the baseline methods on the test sets. Figure. 3 visualizes the influence of adding more landmarks in a scene on localization for one of the test sets. Both plain SFA and PoseNet performance degrades when the test set (i.e., test set number 2) differs considerably from the training set.

2) Large-scale Garden: The training and test set contains images from the trajectory, as shown in the fig. 2d. The training set has 2055 images, while the test sets consist of 1781 and 914 images. After training PoseNet, the localization error on the validation set is 0.17m. The large-scale garden has five pre-trained anchors. Table III reports the detection rates and the median localization performance. The detection rates of the landmarks drop because of their visibility in a specific region of the scene only. In this experiment, the anchor-based localization performs better than the baseline methods for both test sets. However, these anchors are typically dynamic objects (i.e., a suitcase, a bicycle) in a scene. Hence, localization in the long-term would break if these anchors are no longer present or moved to another place in the scene. On the other hand, the learned landmarks are stable regions and thus can be reliably used for localization in the long term. The individual landmarks enable coarse localization performance, and their combination achieves similar localization accuracy to PoseNet in this experiment. The addition of more landmarks can further improve the localization accuracy. The reduced performance for the second test set is due to the significant condition changes from the training set (i.e., sun glare, reflections).

C. Scaling Experiments

This experiment aims to analyze the effect of increasing the number of landmarks on localization using real-world data from the large-scale garden. We used ten landmarks derived using the anchors present in the scene. The first step is to learn an independent position estimator for each landmark. The second step processes landmark images from the test set using the estimators and predict the robot's 2D position (x, y). Afterwards, we calculate the test set's median localization error by systematically increasing the number of landmarks. Fig. 4 shows the results of 50 random permu-

TABLE II

Real-world Small-scale Localization: Median localization performance w.r.t anchors, learned landmarks, and the baseline methods. The learned landmarks achieve similar detection rates as the pre-trained anchors. The individual landmarks enable coarse scene localization while their combination performs similarly or better than the baseline methods.

Test Set	Landmark-based Localization						Image-SFA	PoseNet
		Id_1	Id_2	Id_3	Combined	%		
1	**Anchors**	0.33m [99 %]	0.54m [99 %]	0.27m [99 %]	0.19m	100	0.19m	**0.18m**
	Learned	0.60m [98 %]	0.31m [98 %]	0.26m [97 %]	**0.18m**	100		
2	**Anchors**	2.35m [99 %]	1.39m [98 %]	0.66m [99 %]	0.63m	100	1.01m	0.83m
	Learned	1.33m [99 %]	0.93m [97 %]	0.73m [98 %]	**0.55m**	100		

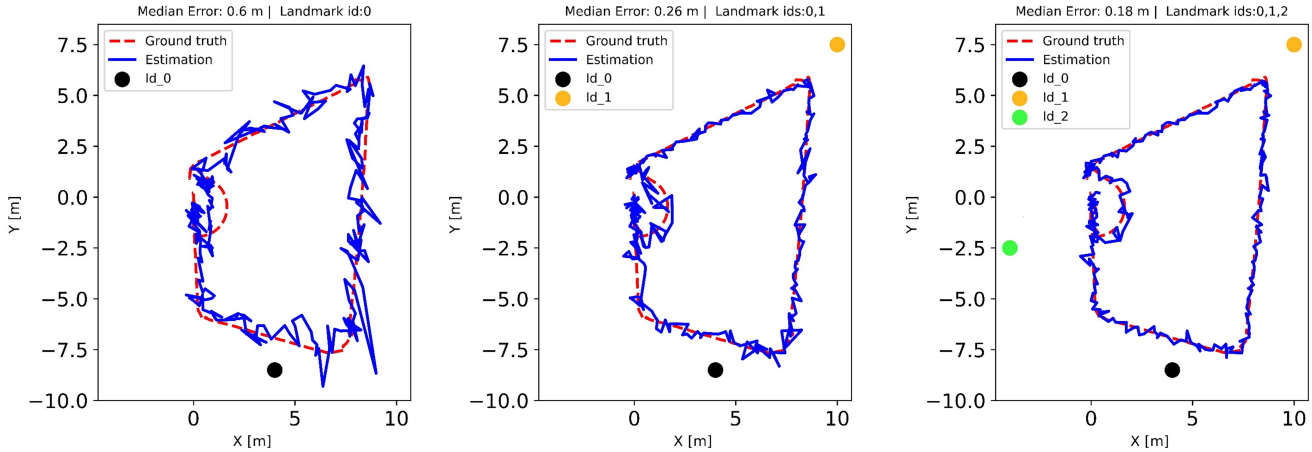

Fig. 3. **Influence of Landmarks on Localization:** A human interactor iteratively adds new anchors in a scene. The robot learns the landmarks relative to the placed anchors for the localization task. Using a single landmark gives a rough estimate of the robot's position (left) while adding more landmarks provides the precise location estimation (right). Hence, the robot can cooperatively scale the system to achieve task-dependent localization accuracy.

TABLE III

Real-world Large-scale Localization: Median localization performance w.r.t anchors, learned landmarks, and the baseline methods. The drop in the detection rates is due to the landmark visibility in a specific part of the scene. The anchor-based localization performs better than the baseline methods. However, they can not be used for long-term localization since they are mobile objects (i.e., a bicycle). Individual landmarks enable coarse localization while their combination performs better than the plain SFA and is similar to PoseNet in both experiments.

Test Set	Landmark-based Localization								Image-SFA	PoseNet
		Id_1	Id_2	Id_3	Id_4	Id_5	Combined	%		
1	**Anchors**	0.60m [27 %]	0.92m [58 %]	1.83m [86 %]	1.46m [60 %]	0.69 [49 %]	**0.55m**	100	3.56m	0.76m
	Learned	0.58m [25 %]	1.17m [57 %]	2.91m [86 %]	2.29m [54 %]	1.09 [45 %]	0.81m	100		
2	**Anchors**	0.99m [28 %]	4.21m [60 %]	3.99m [83 %]	3.54m [38 %]	4.76 [41 %]	**1.91m**	100	6.90m	2.37m
	Learned	1.62m [26 %]	7.02m [60 %]	5.58m [83 %]	3.33m [32 %]	4.88 [37 %]	2.64m	100		

tations of the ten landmarks for both test sets. Both plots show an improved localization performance by increasing the number of landmarks until it saturates as expected. The localization error for the first test set using ten landmarks is 0.42m, which is almost 51% better than the result obtained with five landmarks (c.f. table III). Similarly, it improved by 55% (1.43m) for the second test set with ten landmarks. From an application perspective, this implies that a robot could increase the number of landmarks to achieve a certain accuracy level at runtime.

IV. CONCLUSION

In this work, we proposed a new approach that aims to speed up the label generation process for learning new landmarks. The method uses physical instances of pre-trained CNN objects as anchors to generate labeled data for the unseen imagery. Pretrained detectors like MS-COCO are mainly available for common manufactured mobile objects, which makes them unsuitable as landmarks. Our approach turns this disadvantage into an advantage since common mobile objects can be easily acquired and placed temporarily as anchors. Therefore, a human only needs to place an anchor in the scene and specify its spatial relationship to a new landmark once instead of potentially hand-labeling thousands of images. Please note that a human is not strictly required to suggest new landmarks but to provide a temporarily stable and unique training anchor object. However, a human's common sense knowledge can help select

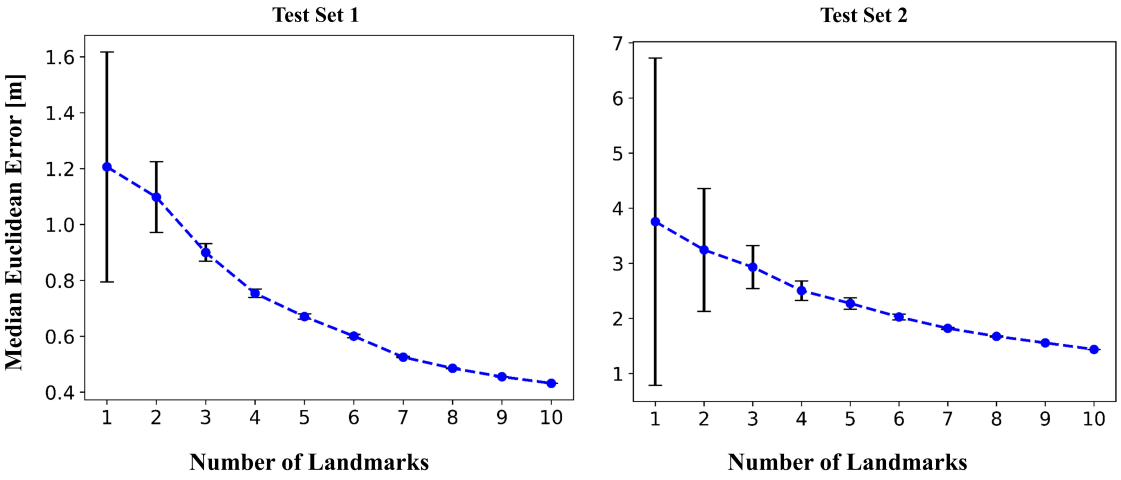

Fig. 4. **Effect of Increasing Landmarks on Localization:** Test set 1 (Left) and test set 2 (Right) from the large-scale garden. The plot shows the median and variance of the localization performance for 50 random permutations of the ten landmarks. The usage of more landmarks for localization improves the performance initially, and eventually, it saturates for a higher number of landmarks.

landmarks that will be stable and unique over time (e.g., manufactured objects rather than vegetation to not change too much with changing seasons). Afterwards, the system automatically generates the labeled data and trains a detector to learn the landmarks. Most machine-learning approaches work on prerecorded labeled image data, whereas our approach requires physical interaction on-site with the robot. After the landmark learning phase, we used the landmark views to perform localization experiments. The results show similar or better localization performance than the baseline methods. The method's performance can be further improved by integrating more landmarks and obtaining a combined position estimation relative to them. Our approach uses SFA as an efficient unsupervised feature learning step. However, we expect physical interactive landmark learning to also improve standard supervised pose regression methods like PoseNet.

The labeling approach allows learning actual semantic objects as new landmarks. However, the new landmark may not necessarily contain complete objects; it can be a non-object patch due to perspective effects. As long as the patch is informative (e.g., not a part of the sky), our results show that localization learning performs well on such landmarks. From an application perspective, our system is suitable for service robots (e.g., lawnmowers and vacuum cleaners), employing a pre-existing visual detector to learn new landmarks in a scene. Thus, the approach enables reliable localization in the long-term even if the anchor objects are no longer present in the scene.

References

[1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.

[2] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-Scale Direct Monocular SLAM," in *Computer Vision - ECCV 2014, Proceedings, Part II*, 2014, pp. 834–849.

[3] A. Kendall, M. Grimes, and R. Cipolla, "Convolutional networks for real-time 6-dof camera relocalization," *CoRR*, 2015.

[4] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," *CoRR*, 2017.

[5] M. J. Milford, G. F. Wyeth, and D. Prasser, "RatSLAM: a hippocampal model for simultaneous localization and mapping," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, vol. 1, April 2004, pp. 403–408 Vol.1.

[6] M. Franzius, H. Sprekeler, and L. Wiskott, "Slowness and Sparseness Lead to Place, Head-Direction, and Spatial-View Cells," *PLoS Computational Biology*, vol. 3, no. 8, pp. 1–18, 2007.

[7] B. Metka, M. Franzius, and U. Bauer-Wersing, "Bio-inspired visual self-localization in real world scenarios using Slow Feature Analysis," *PLOS ONE*, vol. 13, no. 9, pp. 1–18, 09 2018.

[8] D. Frost, V. Prisacariu, and D. Murray, "Recovering Stable Scale in Monocular SLAM Using Object-Supplemented Bundle Adjustment," *IEEE Transactions on Robotics*, vol. 34, no. 3, pp. 736–747, 2018.

[9] E. Sucar and J. Hayet, "Bayesian Scale Estimation for Monocular SLAM Based on Generic Object Detection for Correcting Scale Drift," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 5152–5158.

[10] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *Trans. Rob.*, vol. 28, no. 5, p. 1188–1197, Oct. 2012.

[11] M. Haris, M. Franzius, and U. Bauer-Wersing, "Visual Localization and Mapping with Hybrid SFA," in *Conference on Robot Learning (CoRL)*, 2020.

[12] M. Sarrica, S. Brondi, and L. Fortunati, "How many facets does a "social robot" have? a review of scientific and popular definitions online," *Information Technology and People*, vol. 33, 04 2019.

[13] A. Henschel, G. Laban, and E. Cross, "What makes a robot social? a review of social robots from science fiction to a home or hospital near you," *Current Robotics Reports*, vol. 2, 03 2021.

[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[15] M. Haris, M. Franzius, and U. Bauer-Wersing, "Robust outdoor self-localization in changing environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2019)*. IEEE, 2019.

[16] L. Wiskott and T. Sejnowski, "Slow Feature Analysis: Unsupervised Learning of Invariances," *Neural Computation*, vol. 14, no. 4, pp. 715–770, 2002.

[17] T. Zito, N. Wilbert, L. Wiskott, and P. Berkes, "Modular toolkit for Data Processing (MDP): a Python data processing framework," *Front. Neuroinform.*, vol. 2, no. 8, 2009.