Proceedings of Machine Learning Research 176:349–354, 2022 NeurIPS 2021 Competition and Demonstration Track

Deep Learning Frameworks for Weakly-Supervised Indoor Localization

Farhad G. Zanjani[†], **Ilia Karmanov**[†], **Hanno Ackermann, Daniel Dijkman Simone Merlin, Ishaque Kadampot, Brian Buesker, Vamsi Vegunta Fatih Porikli**

Qualcomm AI Research[*]

{fzanjani,ikarmano,fporikli}@qti.qualcomm.com

Editors: Douwe Kiela, Marco Ciccone, Barbara Caputo

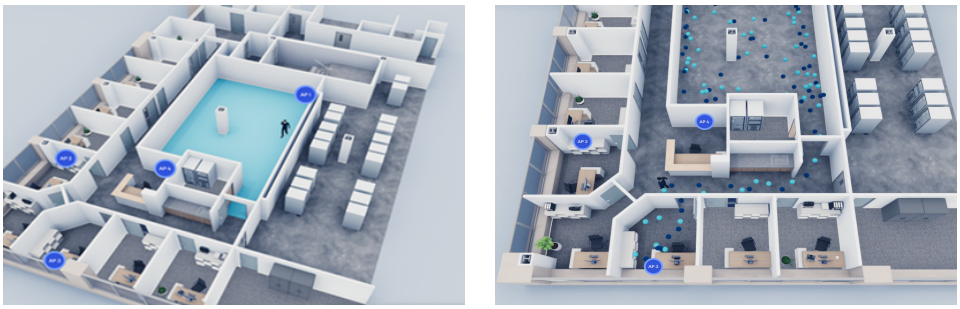


Figure 1: Demonstration environment for person localization via Wi-Fi.

Abstract

We present two weakly-supervised deep learning frameworks for person indoor localization through Wi-Fi signal. These two frameworks, namely OT-Isomap and WiCluster, in contrast with prior works, require only room/zone level labels that is easier to acquire, compared to hard-to-acquire centimeter accuracy position labels. The OT-Isomap is a modality-agnostic model and formulates the localization problem in the context of parametric manifold learning and optimal transportation. This framework allows jointly learning a low-dimensional embedding as well as correspondences with a topological map. The WiCluster method is based on self-supervised deep clustering and metric learning models. Inspired by the deep cluster method, the Wi-Fi signals are spatially charted and represented in lower-dimensional space while a triplet margin-loss constrains an isometric representation of data on its 2D/3D intrinsic space. We demonstrate the meter-level accuracy of these two methods on both real-world Wi-Fi and camera-based indoor localization.

Keywords: Indoor localization, parametric manifold learning, deep metric learning

1. Introduction

Self-localization or localizing objects in the scene are primary tasks in navigation and surveillance systems. This problem has been the subject of many studies in the machine learning community. It has been addressed in several areas such as visual odometry (Engel

[*] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
[†]. Contributed equally

et al., 2017; Brahmbhatt et al., 2018), visual simultaneous mapping and localization (VSLAM) (Davison et al., 2007; Mur-Artal et al., 2015), self-localization (Arth et al., 2011; Sattler et al., 2019; Sarlin et al., 2021), etc., and many approaches have been proposed. This problem has been studied in different fields such as computer vision, wireless sensing, acoustic sensing, robotics, etc. For example, recent localization methods achieve decimeter precision in the positioning of a moving camera in an indoor environment by leveraging the advances in neural networks (Brahmbhatt et al., 2018; Kendall et al., 2015). However, such techniques are highly entangled with the modality of data in use; thus, applying such specialized algorithms to other modalities is often not possible. For example, the existing visual odometry and VSLAM techniques that rely on visual features or camera projection models are incompatible with different sensory systems like radio frequency (RF) or audio signals. Still, all instantiate the same problem. In contrast to existing solutions tailored for a particular modality, our methods involve minimal assumption regarding the data modality in use, hence they can be easily adapted to a new data modalities for localization task.

There are also many classical methods based on digital signal processing for indoor positioning (Qian et al., 2018; Xie et al., 2019) that can determine the location of subjects by assuming that the perturbation of the RF signal propagation, induced by the target motion, follows a known mathematical model. In reality, such models work when the target is within line-of-sight of the transmitter and receiver, but fail in environments with non-line-of-sight propagation and with complex reflection patterns. Moreover, these works mostly are applicable to RF signal and don't generalize to other sensory inputs that may be used for localization.

In this demonstration report, we first explain briefly the OT-Isomap (G. Zanjani et al., 2021) and WiCluster (Karmanov et al., 2021) methods and then we show the evaluation results on real-world Wi-Fi data and public image data. For detailed technical explanation of each method, we refer the reader to the original papers.

2. Weakly-supervised neural frameworks

OT-Isomap (G. Zanjani et al., 2021) assumes the measured samples in input ambient space lie on a smooth manifold; thus, the manifold is locally connected. This assumption intuitively holds since the data is a temporal sequence that does not change much between two consecutive instances in time. It also assumes that the topological map that represents the geometry of the environment is known, yet no correspondences between these two spaces are available. This map can be in the form of a 2D sketch or a floor plan of a building. Localizing the observer on the map requires finding a mapping between input space and the target space. OT-Isomap performs this mapping in two stages that are performed jointly. First it maps the samples from their ambient input space into their intrinsic 2D space. For preserving the pairwise distances and learning an isometric transformation, it uses a neural network e.g. an MLP, while preserving the pairwise distances in the intrinsic space. Since there is no ground truth position labels, for finding the correspondences between 2D intrinsic space and the map of building, it employs the optimal transportation technique. For more details, we refer to the original paper.

WiCluster (Karmanov et al., 2021) leverages the inductive prior in sequential positioning data, that within a small window, a sample closer in the sequence will also be closer in

Table 1: Localization error (in meters) for 15 scenes of iGibson data set using OT-Isomap.

	Environments														
index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
error	0.62	1.05	0.73	0.64	0.69	0.83	1.02	1.03	0.73	0.62	0.74	0.71	0.98	1.04	1.14

position. However, enforcing this with a triplet-margin loss will only provide local structure since it will preserve the nearest neighbours, similar to other methods like t-SNE (et al., 2008). To enforce global structure a self-supervised criterion is added, where cluster membership in the high-dimensional embedding space must be preserved in the low-dimensional projection. This is accomplished by extracting a set of pseudo-labels corresponding to the clusters assigned by the K-Means algorithm to data in the high-dimensional embedding space, projecting down to a 2/3D space with a neural network, and then predicting the assigned pseudo-labels with a cross-entropy loss. Performing this on a batch-level encourages a low-rank prior in the projection. The combination of these self-supervised criteria results in a learnt mapping that preserves isometry. In the weakly-supervised regime a corresponding map of the environment is added, and predictions that fall outside of this are penalised relative to their extent outside of the map. For more details, we refer to the original paper.

3. Demonstrations

Camera-based indoor localization: Our deep learning frameworks have minimal assumption about the data modality in use and are applicable to a diverse sensory inputs. To demonstrate this capability, we setup an experiments with vision sensors for indoor localization of an agent equipped with camera. In this experiments, we applied the OT-Isomap on panorama image sequences of several indoor spaces of *iGibson* dataset (Shen et al., 2020)(see Appendix). The data set consists of indoor 3D scans of 15 buildings that can be explored interactively (Xia et al., 2020). The scans are created from real homes and offices. Figure 2 shows an example of constructed panorama image, the ground truth and predicted trajectories for an environment, using OT-Isomap. Table 1 shows the localization error of OT-Isomap for these 15 environments. More results added to the Appendix.

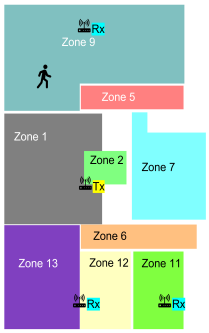
Figure 2: An example of the iGibson environment; (left) the quadruple-view at the location of observer and the constructed panorama image; (middle) a ground truth trajectory; (right) prediction

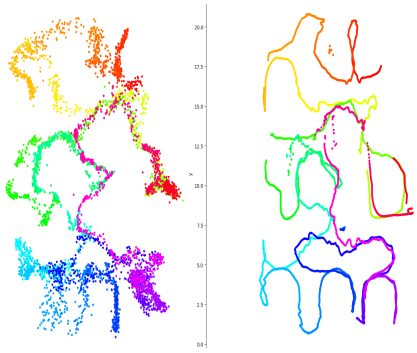
Passive Wi-Fi indoor localization: We setup a real-world deployment with a 2D and 3D data acquisition from two different environments. For the 2D environment of size

15×21 meters in a building, three receivers and one transmitter for a large multi-room space where most areas lack line-of-sight have been used. We collected a data set by using four commercial IEEE 802.11 access points (AP), operating in the 5GHz band. We use three receivers with eight antennas each, and a transmitter with a single antenna. Each receiver collects the *Channel State Information* (CSI) at periodic intervals. The CSI represents the channel between the transmitter antenna and each of its 8 receiver antennas, across 208 frequency tones that span the transmission bandwidth. Hence, the CSI is represented as a multidimensional tensor of complex numbers of dimension $8 \times 1 \times 208$ per each WiFi packet. Table 2 shows the error of WiCluster for person localization in different zones. More experimental results on a different environment and 3D building added to the Appendix.

Table 2: Localization of a person in Wi-Fi signal using WiCluster; (left) mean errors in meters for different zones; (middle) the floor plan image and the defined zones; (right) prediction and ground truth positions.

Zone no.	**WiCluster**		
	Train - walk	***Test - walk***	***Test - run***
1	0.91	0.81	0.74
2	0.82	0.85	0.91
5	0.89	1.40	1.29
6	1.04	1.12	1.31
7	1.99	2.26	2.08
9	1.08	1.29	1.24
11	1.05	1.24	1.28
12	1.29	1.30	1.26
13	1.13	1.05	1.27
average	1.13	1.26	1.26





4. Discussions and Conclusions

We demonstrated two deep learning frameworks that learn the mapping of indoor measured data on the floor plan of a building. Both frameworks learn a 2D/3D embedding space that are isometric and topologically aligned with the given floor plan. Since the proposed frameworks does not use hard-to-acquire position labels, they can facilitate the real-world applications of new emerging sensing data modalities such as Wi-Fi signal for indoor localization.

Acknowledgments

This work was funded by Qualcomm Technologies, Inc.

References

Clemens Arth, Manfred Klopschitz, Gerhard Reitmayr, and Dieter Schmalstieg. Real-time self-localization from panoramic images on mobile devices. In *2011 10th ieee international symposium on mixed and augmented reality*, pages 37–46. IEEE, 2011.

Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2018.

Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007.

Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017.

Maaten et al. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86): 2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Farhad G. Zanjani, Ilia Karmanov, Hanno Ackermann, Daniel Dijkman, Simone Merlin, Max Welling, and Fatih Porikli. Modality-agnostic topology aware localization. *Advances in Neural Information Processing Systems*, 34, 2021.

Ilia Karmanov, Farhad Zanjani, Ishaque Kadampot, Simone Merlin, and Daniel Dijkman. Wicluster: Passive indoor 2d/3d positioning using wifi without precise labels. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–7. IEEE, 2021.

Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015.

Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. Widar2. 0: Passive human tracking with a single wi-fi link. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pages 350–361, 2018.

Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3247–3257, 2021.

Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3302–3312, 2019.

Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín*, Linxi Fan, Guanzhi Wang, Shyamal Buch, Claudia D'Arpino, Sanjana Srivastava, Lyne P Tchapmi, Kent Vainio, Li Fei-Fei, and Silvio Savarese. igibson, a simulation environment for interactive tasks in large realistic scenes. *arXiv preprint arXiv:2012.02924*, 2020.

Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. *IEEE Robotics and Automation Letters*, 5(2):713–720, 2020.

Yaxiong Xie, Jie Xiong, Mo Li, and Kyle Jamieson. md-track: Leveraging multi-dimensionality for passive indoor wi-fi tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019.