Learning Scan Context toward Long-term LiDAR Localization

Giseop Kim[1] and Byungjae Park[2] and Ayoung Kim[1]*

***Abstract*— We present a robust Light Detection and Ranging (LiDAR)-based place descriptor called Scan Context Image (SCI) and a localization method that uses this representation for long-term simultaneous localization and mapping (SLAM). We formulate localization as a conventional supervised classification problem using convolutional neural network (CNN), where a gridded place is considered as a single class and SCIs acquired in that place are the corresponding data to the class. The small (three layers in our work) network is trained with only a single sequence from a single date, which has a total of 579 places and an average 26 data per class. Despite these constraints, we show that a learned network achieves sufficient performance in outdoor (top 5 average accuracy is over 90%) on 13 other test sequences from different dates over 1 year with varying environmental conditions.**

I. Introduction and Related works

Robust place recognition over time or long-term localization is a crucial module of robot navigation in the real world beyond indoor or simulated environments. Among many factors, temporal variance of the environment causes major challenges for long-term localization. To achieve this long-term localization, however, it needs to overcome both appearance (e.g., seasonal or weather changes) and structure (e.g., occlusions, foliage, or constructions) variations. Recently, when dealing with the aforementioned issues, deep learning-based approaches are drawing attention due to their versatile performance under temporal and visual variations.

Deep learning methods for localization can be categorized into two types according to their roles. The first category serves as a robust feature extractor. Sünderhauf et al. [2] showed using only robust regions in an image, whose objectness scores are high, can improve place recognition performance under severe appearance and viewpoint changes. More recently, Naseer et al. [3] proposed a lighter method that leaves a continuous robust region, not bounding boxes, in an image via learning up-convolutional networks such as [4, 5]. However, the methods in the first category require additional matching procedures after extracting robust features. The second category supports end-to-end localization. The output of this type is a metric or topological pose. PoseNet [6] is a seminal work; it regresses the 6D pose of a query image using CNN. Walch et al. [7] extended [6] by combining long short-term memory (LSTM) units, reporting a better performance than PoseNet. PlaNet [8] tried to address the topological localization in a gridded map using a classification network.

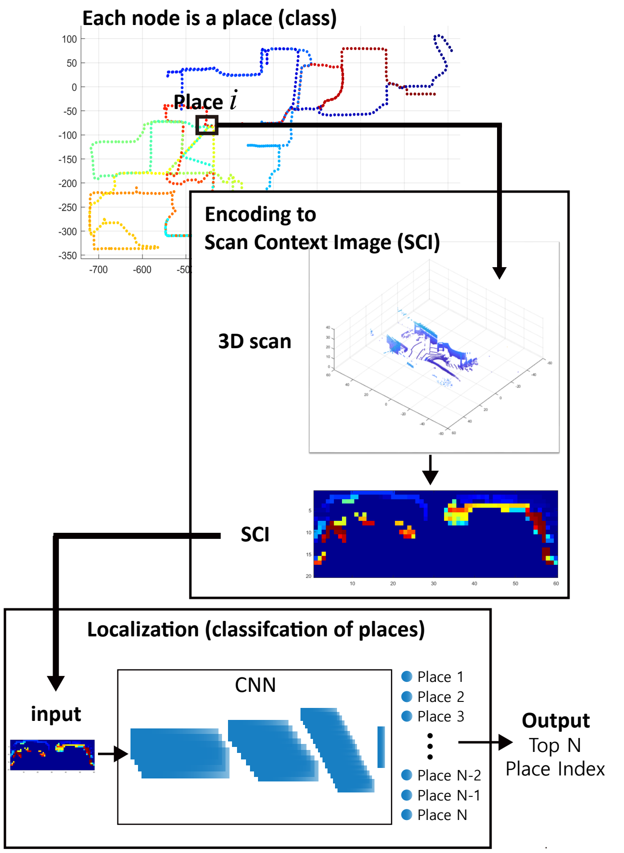


Fig. 1. Overview of the proposed long-term topological localization system. Each equi-distant place (as in Fig. 3) is considered a unique class. A 3D scan acquired in a place is encoded into the SCI improved from [1] and the SCI is fed to the CNN whose result is a classification probability vector of places. The network was trained using only one day in this paper. Because the SCI effectively summarizes the ego-centric structural information of the environment, it recognizes a place with high accuracy even on different dates and in different environmental conditions over a year.

As described above, many long-term place recognition or localization studies have been conducted on applying deep learning to visual sensor data, but there are few studies using LiDAR sensor data with deep learning for robust long-term localization. The data obtained from LiDAR such as a point cloud can inherently avoid the difficulties that occur from image-based localization because it directly captures the structural information of space, which is independent of the illumination variance, and because structural information often changes in a slower rate than visual

[1] G. Kim and A. Kim are with the Department of Civil and Environmental Engineering, KAIST, Daejeon, S. Korea `[paulgkim, ayoungk]@kaist.ac.kr`

[2] Byungjae Park is with the Intelligent Robot System Research Group, ETRI, 218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, Republic of Korea. `bjp@etri.re.kr`

appearance. Thus, adopting LiDAR sensor data (e.g., point cloud) could be more promising for long-term localization. Furthermore, city-scale point cloud maps are nowadays being easily generated [9]. Despite their availability, however, there are few studies that effectively utilize them as preliminary information for long-term localization of robots due to the size of the point cloud maps.

In this paper, we propose a compact representation that summarizes a place within a 3D point cloud and strategies for learning the representation using CNN. In our previous work [1], we showed that the proposed representation, called *Scan Context*, is efficient and effective for online place recognition. Scan context directly summarizes a 2.5D shape of a visible scene from the observer at the location using a 3D LiDAR scan. The accompanied matching algorithm provides a fast coarse yaw alignment (in that work, 6° resolution) between 3D scans so that the scan context is robust to view-point changes (e.g., corner or reverse loop) and serves as a good initial value for finer localization (e.g., Iterated Closest Point (ICP)).

In this paper, we demonstrate that scan context is also effective for long-term robot localization as well as place recognition. We formulate the topological localization problem as a classification of places and train the CNN by learning scan contexts as in Fig. 1. The contributions of the paper are summarized as below.

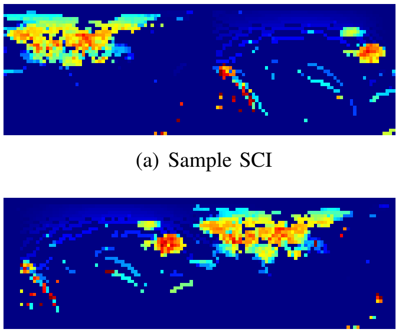
- We apply structure-induced synthetic images for deep learning based localization. Thereby, we exploit deep learning while conveying urban structures.
- We propose data preprocessing and augmentation methods for effectively training a network to increase its performances.
- We empirically prove that learning a small (three layers in this work) network with few training scan contexts (e.g., a single trajectory from a single day) guarantees consistent and adequate performances on the other dates, whose environmental conditions vary over a year.

II. Methodology

In this paper, we formulate the topological localization problem as a conventional classification problem using CNN, such as in PlaNet [8]. This means that a single location (e.g., $10\,\mathrm{m}$ by $10\,\mathrm{m}$ grid cell) is treated as a single class. Therefore, the localization is equivalent to predicting a label of a query SCI. First, we briefly introduce the SCI developed in our previous work [1] as an input of a network and then propose training strategies for achieving sufficient performance with a small amount of data and a small network.

A. Scan Context Image (SCI)

Scan context is an ego-centric spatial descriptor generated from a 3D point cloud. The scan context is represented in an image format by mapping azimuthal and radial direction information to the two image coordinates, columns and rows, respectively. We define this synthesized image as SCI. Based on an observer, the height of the structures along the radial direction comprises a column each in the synthesized SCI. Then, evaluating this height change along the radial direction for each azimuth angle yields the final SCI. The resolution of the SCI is determined by the spatial resolution. In this paper, an image of 40 pixels by 120 pixels is synthesized. That is, we used 40 bins ($2\,\mathrm{m}$ resolution when maximum sensing range is set to $80\,\mathrm{m}$) in the radial direction and 120 bins ($3\ ^\circ$ resolution) in the azimuthal direction. This structural description in an image format well discriminates locations from a map. Because the descriptor preserves direction information (column order), rotation invariant detection is achievable via a proper distance function such as [1].



(a) Sample SCI

(b) Augmented SCI (reversed)

Fig. 2. An example of the reversed route augmentation. Shifting half of the columns (180°) acheives data augmentation. This is expected to mimic the scan context that is obtained when visiting the same place from the reverse direction. Both the original and its reversed route SCI with the same label are used for training a network.

B. Preprocessing and Data Augmentation

To better train a network and to increase the discriminative power, raw scan context processing strategies are proposed.

Coloring (3-channel amplification): First, we feed a colorized scan context into a network for both training and test. The original scan context is a 1-channel matrix since a pixel value of a scan context is a single real value, highest height in the corresponding bin. However, we amplify a raw 1-channel scan context to a 3-channel colorized SCI with a specific colormap to improve the discriminative power. In this work, we use jet colormap, and the color axis is set from $0\,\mathrm{m}$ to $15\,\mathrm{m}$. The threshold range was chosen empirically to minimize the skewness of the distribution of height values in the region of the NCLT dataset. That is, the color is blue if the bin value is 0 (e.g., ground) and the color is red if the bin value is equal to or greater than 15 (e.g., a tree or a building).

Backward data augmentation: Our ultimate goal is to learn the SCI for a location from a single day and recognize places on different dates. If we were to exploit a single sequence training, a place is usually visited in only one direction during the training sequence. To variate and impose rotation invariance, we performed data augmentation, as in Fig. 2, to be able to recognize when visiting the same place in a reversed route.

TABLE I

THE NETWORK STRUCTURE AND PARAMETERS

Input	Conv			Max Pool	BN	Conv			Max Pool	BN	Conv			Max Pool	Flat	Drop out	Fully Connected	Drop out	Fully Connected
40x120	Kernel Size	# of Filters	Activ. Func.	Pool Size		Kernel Size	# of Filters	Activ. Func.	Pool Size		Kernel Size	# of Filters	Activ. Func.	Pool Size			# of Units		# of Units
	5x5	64	ReLU	2x2		5x5	128	ReLU	2x2		5x5	256	ReLU	2x2		0.7	64	0.7	914 (# of Classes)

TABLE II

INFORMATION ABOUT TRAINING AND TEST SEQUENCES

Type	Date	Conditions			Length	# of
		Time	Foliage	Snow	(km)	Sampled Data
Train	2012-01-15	Afternoon	No	Yes	7.5	15078
Test	2012-02-04	Afternoon	No	No	5.5	5170
	2012-02-05	Morning	No	No	6.5	6238
	2012-02-12	Midday	No	Yes	5.8	5396
	2012-02-19	Midday	No	No	6.2	5782
	2012-03-17	Morning	No	No	5.8	5449
	2012-03-25	Midday	No	No	5.8	5478
	2012-05-26	Evening	Yes	No	6.3	5536
	2012-06-15	Morning	Yes	No	4.1	3321
	2012-08-20	Evening	Yes	No	6.0	5148
	2012-09-28	Evening	Yes	No	5.6	4624
	2012-11-16	Evening	No	No	4.8	3575
	2013-02-23	Afternoon	No	Yes	5.2	4119
	2013-04-05	Afternoon	No	Yes	4.5	3342

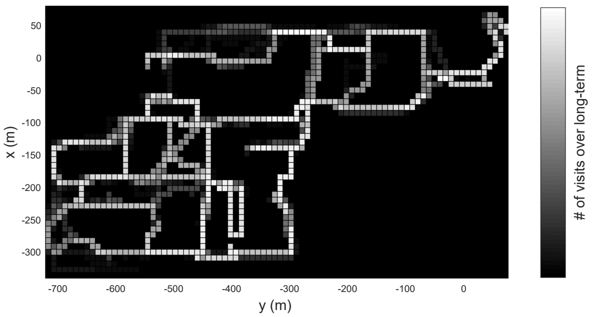


Fig. 3. Places of NCLT area. A place is a 10 m by 10 m grid.

III. EXPERIMENTS AND RESULTS

A. Training the Network

We validate our method using The North Campus Long-Term (NCLT) dataset [10]. This dataset contains long-term multiple sequences that cover similar trajectories with wide variations of environmental conditions (e.g., time, foliage, and weather). The sequences are named by its acquisition date, such as `2012-01-15`.

We first divide the whole area of the NCLT dataset into 10 m by 10 m grids, and each grid is considered as a single place. Fig. 3 shows the gridded places and the frequency of visits over time per each place. The white indicates that the place had been visited more frequently. The NCLT dataset has a total of 914 places with at least one measurement. We numbered each place from 1 to 914 and the index is used as the label of the place.

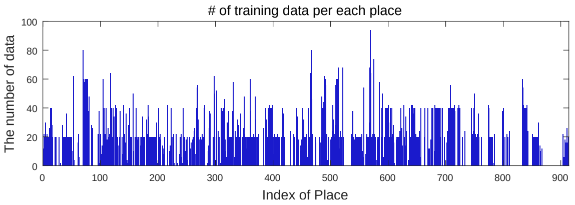


Fig. 4. The distribution of the number of training data per place. A single sequence, `2012-01-15`, is used to train the network. The sequence `2012-01-15` visits 579 places, and each place has 26 data on average.

The architecture of the network and parameters we use are summarized in the Table I. This network takes a 40x120 RGB image (colorized SCI) and classifies the label of the place, where a query SCI is acquired. The network is trained using only a single sequence `2012-01-15`. The scans of the training sequence are sampled at every 1 m. Each scan is encoded into a SCI and is labeled according to its corresponding place index. Fig. 4 show the distribution of the number of training data along each place. Places that are visited multiple times have slightly more data. The label of a place is turned into a 914 dimension one-hot encoded vector to train the classification network. `2012-01-15` has 579 places but we set the dimension of the one-hot vector to 914, which is the total number of places the NCLT dataset has, considering later expandability. We used an Adam [11] optimizer with default parameters from Keras [12]. The batch of 64 is used.

B. Performance Evaluation

We validate the performance of the trained network on 13 test sequences from other dates. Those sequences were selected considering the diversity of environmental conditions and the detailed information in Table II. The scans of the test sequences are also sampled at an equi-interval of 1 m. Each scan is an encoded colorized SCI. In the case of a test, backward data augmentation is not applied. That is, only the scan context acquired in that current direction is fed to the trained network.

Only places visited in the training sequence are manually selected and tested in this preliminary work. The performance is quantitatively measured via (1):

$$\text{Accuracy} = \frac{|\text{SC}^{\text{test}}_{correct}|}{|\text{SC}^{\text{test}}|} \tag{1}$$

, where SC^{test} is a set of all test scan contexts of a sequence and $\text{SC}^{\text{test}}_{correct}$ is a set of scan contexts whose place prediction

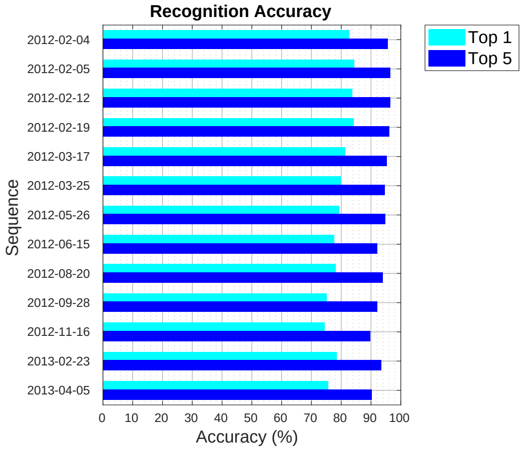


Fig. 5. Classification results. If the correct label is included in the labels having N largest values in the 914 dimension output vector, it is considered true-positive localization.

TABLE III

AVERAGE ACCURACIES WITH VARIANCE ($1\ \sigma$) FOR TOP 1 AND TOP 5.

	Top 1	Top 5
Accuracy (%)	79.63 ± 3.48	93.94 ± 2.26

result is correct within an allowed degree such as top 1 or top 5.

Fig. 5 shows the localization results for each test date. The trained network shows consistent performance with little variance over time. Because SCI summarizes the structural information, the performance of the network is consistent regardless of seasonal, light condition (time), and structural (e.g., foliage and snow) changes. In addition, unlike range images from 3D LiDAR, SCI is also robust against temporal occlusion by summarizing maximum height distribution of overall visible scene structure. The average performances with variances are summarized in Table III. The network retrieved the top 5 places with over 90% accuracy for large-scale outdoor environments (ground areas of at least $50,000\ \mathrm{m}^2$ because each place is a $10\ \mathrm{m}$ by $10\ \mathrm{m}$ grid and the training sequence has 579 places). Thus, it can function as an effective global localizer for scalable localization over long-term period, which means that the SCI network can be used as a forward module to reduce the search space for existing map-based finer localization for metrically accurate localization.

IV. Conclusion

In this paper, we demonstrated the Scan Context is effective for long-term robot localization. We only had a few training samples per place, which is considered as a class, from a single date and showed it guarantees consistent and sufficient performance (top 5 accuracy is over 90% for 579 different places) on other 13 dates over 1 year.

In the future, we plan to compare ours with other ConvNet methods such as [13], and integrate unknown-unknown place detection module within our localization system.

Acknowledgment

This material is based upon work supported by the MOTIE, Korea (No.10051867), IITP grant funded by MSIT (No.2017-0-00067), and [High-Definition Map Based Precise Vehicle Localization Using Cameras and LIDARs] project funded by Naver Labs Corporation.

References

[1] G. Kim and A. Kim, “Scan Context: Egocentric Spatial Descriptor for Place Recognition within 3D Point Cloud Map,” in *Proc. IEEE/RSJ Intl. Conf. on Intell. Robots and Sys.* IEEE, 2018, (Under Review).

[2] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, “Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free,” *Proceedings of Robotics: Science and Systems XII*, 2015.

[3] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, “Semantics-aware visual localization under challenging perceptual conditions,” in *Proc. IEEE Intl. Conf. on Robot. and Automat.* IEEE, 2017.

[4] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. on Comput. Vision and Pattern Recog.*, 2015.

[5] G. L. Oliveira, W. Burgard, and T. Brox, “Efficient deep models for monocular road segmentation,” in *Proc. IEEE/RSJ Intl. Conf. on Intell. Robots and Sys.* IEEE, 2016.

[6] A. Kendall, M. Grimes, and R. Cipolla, “Posenet: A convolutional network for real-time 6-dof camera relocalization,” in *Proc. IEEE Intl. Conf. on Comput. Vision*. IEEE, 2015.

[7] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, “Image-based localization using LSTMs for structured feature correlation,” in *Proc. IEEE Intl. Conf. on Comput. Vision*, 2017.

[8] T. Weyand, I. Kostrikov, and J. Philbin, “PlaNet - Photo Geolocation with Convolutional Neural Networks,” in *Proc. European Conf. on Comput. Vision.* Springer, 2016, pp. 37–55.

[9] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, “Complex Urban LiDAR Data Set for high Resolution 3D Environment,” in *Proc. IEEE Intl. Conf. on Robot. and Automat.* IEEE, 2018.

[10] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, “University of Michigan North Campus long-term vision and Lidar dataset,” *Intl. J. of Robot. Research*, 2015.

[11] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv:1412.6980*, 2014.

[12] F. Chollet *et al.*, “Keras,” https://keras.io, 2015.

[13] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “NetVLAD: CNN architecture for weakly supervised place recognition,” in *Proc. IEEE Conf. on Comput. Vision and Pattern Recog.*, 2016, pp. 5297–5307.