Compressive Self-localization Using Relative Attribute Embedding

Yamamoto Ryogo   Tanaka Kanji

***Abstract*— The use of relative attribute (e.g., beautiful, safe, convenient) -based image embeddings in visual place recognition, as a domain-adaptive compact image descriptor that is orthogonal to the typical approach of absolute attribute (e.g., color, shape, texture) -based image embeddings, is explored in this paper.**

I. Introduction

Most current state-of-the-art visual place recognition (VPR) algorithms employ absolute attribute (e.g., color, shape, texture) -based image embedding for image feature description [1]–[3] and image similarity search [4]. In this study, we are interested in relative attributes (e.g., beautiful, safe, convenient) [5]–[7]-based image embedding, as it provides a domain-adaptive ranking-based image description [8] and it is orthogonal to typical approach of absolute attributes-based embeddings. Specifically, we present two different solutions based on binary and real-valued relative attribute strength and experimentally evaluate them via cross-season VPR experiments [9].

II. Approach

VPR is formulated as a problem of similar image retrieval [10]–[12]. The objective is to search for the image most relevant to a given query image over an image database. The database is constructed as a collection of viewpoint-annotated view images from visual experiences in the training domain via structure-from-motion [13] or SLAM [14]. Specifically, the procedure for construction consists of two steps (Fig. 1): (1) extracting a feature descriptor from the image, and (2) evaluating the descriptor similarity between the query and each database images. Either step is detailed in the following.

A. Feature Descriptor

An input query/database image is described by measuring the relative attribute strength in $\mathbb{R}$, with respect to a predefined prototype image [8]. If the strength value is negative, it means that the input image has stronger relative attribute than the prototype image, otherwise it means that it is weaker. (1) Specifically, the processing begins by evaluating the relative attribute strength $p_i^{\textbf{attr}}$ of each i-th prototype image p_i with respect to the input image q for each j-th relative attribute model A_j, which yields a length $(N+1)$ list of relative attribute strength

$$L = (q_j^{\textbf{attr}}, p_{j1}^{\textbf{attr}}, \cdots, p_{jN}^{\textbf{attr}}) \quad (1)$$

Our work has been supported in part by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) 17K00361 and 20K12008.

The authors are with Graduate School of Engineering, University of Fukui, Japan. tnkknj@u-fukui.ac.jp

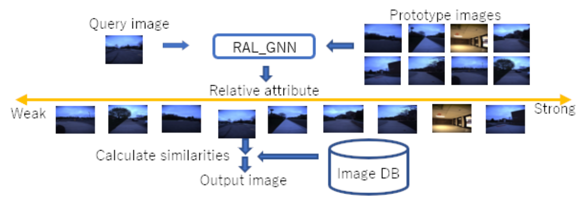

Fig. 1. Relative attribute-based embedding and its application to VPR.

with a boundary condition

$$q_j^{\textbf{attr}} = 0. \quad (2)$$

(2) Then, the list is sorted in the descending order of relative attribute strength, which yields an ordered list L'. (3) Then, each image is ranked based on the ranking L', which yields a $1\times(N+1)$ matrix R_j of rank:

$$(q_j^{\textbf{rank}}, p_{j1}^{\textbf{rank}}, \cdots, p_{jN}^{\textbf{rank}}). \quad (3)$$

(4) By performing the above processes for a predefined set of M relative attributes, we obtain an $M\times(N+1)$ matrix R as the final output of the feature descriptor step.

B. Descriptor Similarity

Next, descriptor similarity is evaluated between the input descriptor R and each database descriptor R', for which we have developed two different kinds of evaluation methods. The first method, called binary relative strength (BRS), treats the descriptor as a binary relative attribute (stronger or weaker), and evaluates the similarity by

$$BRS = \sum_j |R_{j1} - R'_{j1}|. \quad (4)$$

The second method, called ranked relative strength (RRS), utilizes the real-valued strength from the descriptor, and evaluates the similarity by

$$RRS = \sum_i \sum_j |R_{ji} - R'_{ji}|. \quad (5)$$

III. EXPERIMENTS

The experimental settings follow the procedure in [15]. The NCLT dataset used is a large scale, long-term autonomy dataset collected by a Segway robot in a university campus. Specifically, view images from the on-board front-facing camera (Ladybug camera) in the sessions 2012/03/31 and 2012/08/04 were used as training and test image sets, respectively. We considered a place classification task with a set of 8 place classes. Specifically, the entire workspace

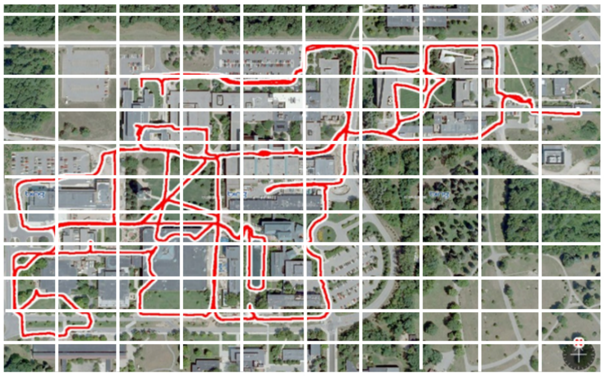

Fig. 2. The robot workspace and place classes.

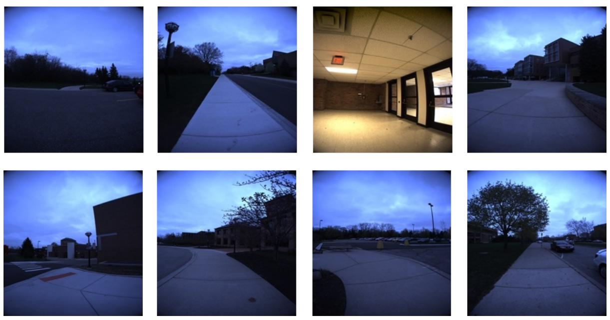

Fig. 3. Prototype images.

of the range $[-740, 130] \times [-330, 120]$ was divided via grid-based place partitioning into a 10×10 grid of 100 place classes (Fig. 2), from which the 8 classes are randomly sampled. We simply sample $N = 8$ prototype images from each of the 8 place classes (Fig. 3). The relative attribute models were trained using the OSR dataset as in [16] (Fig. 4) on the $M = 6$ different attribute classes: natural, open, perspective, large-objects, diagonal-plane and close-depth. The training, testing and performance evaluation were iterated for 100 sets of randomly sampled 8 place classes. The mAP performance was 0.341 ± 0.071 and 0.364 ± 0.079 for the BRS and RRS methods, respectively. For comparison, absolute attribute counterparts of BRS and RRS were also developed using the 1-hot semantic histogram as the absolute attribute feature as in [15], and tested with results of 0.147 ± 0.068 and 0.310 ± 0.090, respectively. One can see that the RRS method outperforms the BRS method in the current experiments. Specifically, the real-valued relative attribute strength provided rich information, and the information loss was significant when it was binarized (i.e., the BRS method).

IV. Conclusions and Future Workds

The use of relative attribute (e.g., beautiful, safe, convenient) -based image embeddings in visual place recognition, as a domain-adaptive compact image descriptor that is orthogonal to the typical approach of absolute attribute (e.g., color, shape, texture) -based image embeddings, is explored in this paper. In the future, we plan to integrate the proposed highly-efficient VPR method to train a visual navigation

Fig. 4. Example images of 8 different categories in the OSR dataset.

system as in [15]. During the training phase, the VPR module must be repeated a very large number of times. That is, the robot needs to experience a large number of (e.g., tens of thousands of) training episodes, and each training episode involves performing VPR at many (e.g., 10) viewpoints. Towards that goal, further acceleration of the VPR module while retraining the discriminative power is desired.

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, September 28 - October 2, 2015*, 2015, pp. 4297–4304.

[3] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[4] M. Cummins and P. M. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *I. J. Robotics Res.*, vol. 30, no. 9, pp. 1100–1123, 2011.

[5] Y. Souri, E. Noury, and E. Adeli, "Deep relative attributes," in *Asian conference on computer vision*. Springer, 2016, pp. 118–133.

[6] R. N. Sandeep, Y. Verma, and C. Jawahar, "Relative parts: Distinctive parts for learning relative attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3614–3621.

[7] D. Parikh and K. Grauman, "Relative attributes," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 503–510.

[8] K. Tanaka, "Deep simbad: Active landmark-based self-localization using ranking-based scene descriptor," *arXiv preprint arXiv:2109.02786*, 2021.

[9] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of michigan north campus long-term vision and lidar dataset," *The International Journal of Robotics Research*, vol. 35, no. 9, pp. 1023–1035, 2016.

[10] S. M. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. D. Cox, P. I. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[11] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognition*, vol. 113, p. 107760, 2021.

[12] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" *arXiv preprint arXiv:2103.06443*, 2021.

[13] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[14] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[15] M. Yoshida, R. Yamamoto, and K. Tanaka, "S3G-ARM: highly compressive visual self-localization from sequential semantic scene graph using absolute and relative measurements," *CoRR*, 2021.

[16] Z. Meng, N. Adluru, H. J. Kim, G. Fung, and V. Singh, "Efficient relative attribute learning using graph neural networks," in *ECCV*, 2018, pp. 552–567.