Data-efficient representation learning for visual place recognition

Leyva Vallina, María

DOI:
10.33612/diss.736449452

Summary

This thesis investigates the problem of visual place recognition, which is a fundamental part of many visual-based localization systems, and therefore of high value to the computer vision community. In particular, we address two problems in the field: the first half of this dissertation is devoted to the presentation and evaluation of a novel dataset in a very underexplored type of environment, namely garden environments. The second half of this thesis addresses the algorithmic part of visual place recognition and proposes a shift of paradigm to learn visual descriptors that encode stronger, reliable, and quantifiable representations of image similarity. The contributions of this dissertation are as follows.

- In Chapter 2 we introduce the TB-Places dataset, which contains approximately 60k images taken in two garden environments. With this dataset we fill the gap of the lack of data for visual localization in these kinds of environments, which present very particular challenges due to viewpoint, weather and illumination variations in the images, as well as the presence of repetitive, very similar textures that make the detection of discriminative visual cues very challenging. Additionally, we evaluate existing off-the-shelf methods for visual place recognition and analize the obtained results.
- We expand the TB-Places dataset with 8k additional images, as we present in Chapter 3, and subsequently expand the benchmark evaluation, including more methods, and models finetuned specifically for the task. The results that we obtained suggested that there was a lot of room for improvement and that the existing visual place recognition approaches had some fundamental weaknesses.
- In the second half of this dissertation we address one particular weakness that we detected during our work in the first two chapters: existing visual place recognition algorithms learn representations by defining a binary ground truth of visually similar or dissimilar image pairs, and then optimizing a loss function that is essentially meant to discriminate between two well-defined and separable classes. However, visual place

recognition descriptors are supposed to encode representations that give an adequate measure of image similarity when feeding two image representation to a given scoring function (i.e. the Euclidean distance). This is not a binary classification problem, as image similarity is not discrete, but continuous: two images can be completely similar, or completely dissimilar, but also anything in between. From this insight, we re-define the binary ground truth of existing visual place recognition datasets (namely MSLS, TB-Places and 7Scenes) to encode a continuous measure of image similarity calculated from geometrical information. Subsequently, we use this new ground truth to optimize a novel Generalized Contrastive Loss, which instead of being defined only for the extreme cases of the ground truth (0 and 1), takes into account any value $\in [0, 1]$. We present our new method and results in Chapter 4, and we demonstrate that with a straightforward modification, we can simplify the learning pipeline, reduce the training time and outperform the state-of-the-art methods by learning more robust descriptors.

- The Generalized Contrastive Loss, although leading to excellent results, still introduces an artificial binarization to the problem, as the function consists of two terms: the first one pushes together the descriptors of similar images, and the second one pulls apart those of dissimilar image pairs. This can lead to performance issues, as we discovered when training transformer backbones. We thus redefine the Visual Place Recognition entirely and approach it as a regression problem: we just regress an image similarity measure (encoded with the Euclidean distance between two representations) and match it to our annotated similarity ground truth. As we thoroughly present in Chapter 5, this paradigm shift allows us to train transformer backbones in a short time, and obtain better results than when using the GCL function. Moreover, we demonstrate that methods trained using regression reach competitive performance in much fewer training iterations than their equivalents trained to optimize a GCL, demonstrating a very data-efficient approach.