Data-efficient representation learning for visual place recognition
Leyva Vallina, María

DOI:
10.33612/diss.736449452

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2023

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 1

Introduction

Similarity is a measure of how alike or comparable two things are in terms of their characteristics, features, properties, or qualities. It is a concept that is widely used in various fields, including mathematics, statistics, psychology, linguistics and computer science.

In computer science, similarity refers to the degree of resemblance between two or more objects, such as text documents, videos, or images. It is a fundamental concept that is used in several applications, such as information retrieval, data mining, natural language processing and computer vision. Although similarity is a somewhat subjective concept, it can be approximately quantified using different methods, such as correlation coefficients, probability distributions, or distance measures. The choice of method depends on the context and the nature of the data being compared. For instance, the Jaccard similarity (the intersection over union between two sets) and the Levenstein distance (the number of operations needed to transform one sequence into another) are widely used metrics in the field of natural language processing, as they are suitable for measuring the resemblance between pieces of textual information. In computer vision, the Cosine similarity (the cosine of the angle between two vectors) and the Euclidean distance (the L2 distance between two points in a multidimensional space) are widely used to compare image representations.

Within computer vision, visual place recognition is an application in which the measurement of image similarity plays a crucial role. Visual place recognition is the task of recognizing a previously visited location from a new image taken at that location. To achieve that, one has to estimate a degree of similarity between that new image and all the images in a database of previously visited places. This task is not trivial and presents many challenges, such as changes in illumination (night vs day), shifts in viewpoint, seasonal variations, partial occlusions and dynamic scenes (Lowry et al. 2016). Some of these challenges are inherent to the environment for which the task is to be performed. Indoor visual place recognition is not affected, for instance, by seasonal and illumination changes, but it usually involves drastic changes in viewpoint and occlusions (Shotton et al. 2013). Visual place recognition in urban environments is affected by dynamic changes in the
1. Introduction

Figure 1.1: Example images within the TB-Places dataset. (a) shows a reference image and (b), (c), (d) and (e) are images with different annotated degrees of similarity. In (f) we show a 2D position map of the shown images in the reference system of the garden.

scene (i.e. presence of moving cars and pedestrians), and variations of illumination, but since these applications usually involve the use of a car-mounted camera, the changes of viewpoint are limited (Arandjelovic et al. (2016), Torii et al. (2013)). Sub-urban and natural environments are evidently more affected by seasonal and weather changes (Milford and Wyeth (2012), Sunderhauf et al. (2013)). Visual place recognition in gardens is a special case due to its unique environmental characteristics. Gardens are subject to changing weather conditions, seasonal changes, and lighting variations. This variability can make it difficult for visual place recognition algorithms to recognize and differentiate between different places in the garden. Additionally, images taken in garden environments usually have repetitive textures such as grass or bushes, from which it is very difficult to extract meaningful and distinctive features. Moreover, despite the rise in popularity of AI and robotics for gardening, public datasets for garden environments are scarce, and models and algorithms trained in other data do not generalize well to these environments due to the aforementioned particular challenges. In this dissertation, we mind this gap by introducing TB-Places (Leyva-Vallina et al. 2019, LeyvaVallina et al. 2019), a novel place recognition dataset for garden environments with images taken in two test gardens of the H2020 Trimbot project (Strisciuglio et al. 2018). Each image has an associated 6DOF pose and a ground-truth similarity value w.r.t. each possible reference image. We show some examples of the TB-Places dataset in Figure 1.1. We evaluate existing visual place recognition methods on this dataset and the results
that we obtain indicate a lot of room for improvement, as current methods do not generalize well to this kind of environment.

We identified one crucial weakness in existing visual place recognition methods: they are trained to optimize either a triplet or a contrastive loss, relying on binary ground truth. That is, they learn to discriminate between similar and dissimilar pairs of images by essentially clustering image pairs into two classes. As we introduced above, image similarity is a continuous attribute, and represents the degree of commonality between two images, which is quantified by a distance function. When a representation is learned in such a way, where the ground truth only includes the extreme ends of the similarity range (totally similar or totally dissimilar), the learned descriptors will not adequately map the values in between and their Euclidean distance will not be representative of the actual image similarity. In order to solve this issue we automatically re-annotated the MSLS (Warburg et al. (2020)), TB-Places (Leyva-Vallina et al. (2019)) and 7Scenes (Shotton et al. (2013)) datasets using available geometrical metadata to obtain an approximated graded ground truth that includes information of partial similarity among image pairs. We used this ground truth to optimize a novel Generalized Contrastive Loss, a re-definition of the contrastive loss that is defined for ground truth values \( \in [0, 1] \), thus enabling the trained model to explicitly learn measures of partial similarity.

The contrastive learning approach still introduces an artificial binarization to the problem, as the contrastive loss is composed of two terms, one that pushes similar representations together, and another that pulls the representations of dissimilar image pairs apart. This has limitations when training transformer backbones, as we demonstrate in Chapter 5. We completely shift the paradigm for visual place recognition and approach it as a regression problem. We obtain even better results than with the GCL function, tune fewer hyperparameters, and we reach performances on-par with state-of-the-art methods in fewer training iterations. Moreover, when we optimize a regression loss, we can successfully train transformer encoders, which get stuck in sub-optimal minima when we optimize a GCL.

In this dissertation, we explore a shift of approach to visual place recognition by using an inductive bias with graded similarity and regression, instead of a contrastive optimization. This allows us to reach state-of-the-art performance while simplifying the pipeline in many ways.

1.1 Thesis Organization

This dissertation is organized as follows.

1. In Chapter 2 we introduce the TB-Places dataset, a novel benchmark for vi-
1. Introduction

Visual place recognition and image localization in garden environments. The dataset was collected within the context of the Trimbot2020 project [Strisciuglio et al. (2018)], and contains images taken in two gardens over the course of two years. Each image has an associated 6DOF camera pose, and we computed a pairwise ground truth for similar and dissimilar images (i.e. depicting the same place or not).

2. We expand this work in Chapter 3 where we introduce a new curated version of the TB-Places dataset with a new set of images taken a year later in the main Trimbot garden. Moreover, we train and evaluate a siamese convolutional model, for which the obtained results suggested the need for a re-evaluation of the approach to the problem.

3. In Chapter 4 we introduce a shift in the paradigm of visual place recognition, and redefine the concept of image similarity from binary to continuous, and optimize a Generalized Contrastive Loss function that can learn more robust representations for visual place recognition in different environments. We demonstrate that our GCL allows training big convolutional backbones without complex hard-mining strategies, getting better performance with less training time.

4. However, contrastive learning still introduces an artificial binarization to the learning of image similarity by dividing the cost function into a positive and a negative term. We tackle this in Chapter 5 where we approach the VPR problem as a regression, and we demonstrate that this approach obtains even better results with an even simpler pipeline, and can be used to effectively train transformer architectures, something which was not accomplished with the GCL.

Chapters 2, 3 and 4 are published works, while Chapter 5 is currently under review.