A new approach for map densification in visual place recognition

May 22 2023, by Ingrid Fadelli

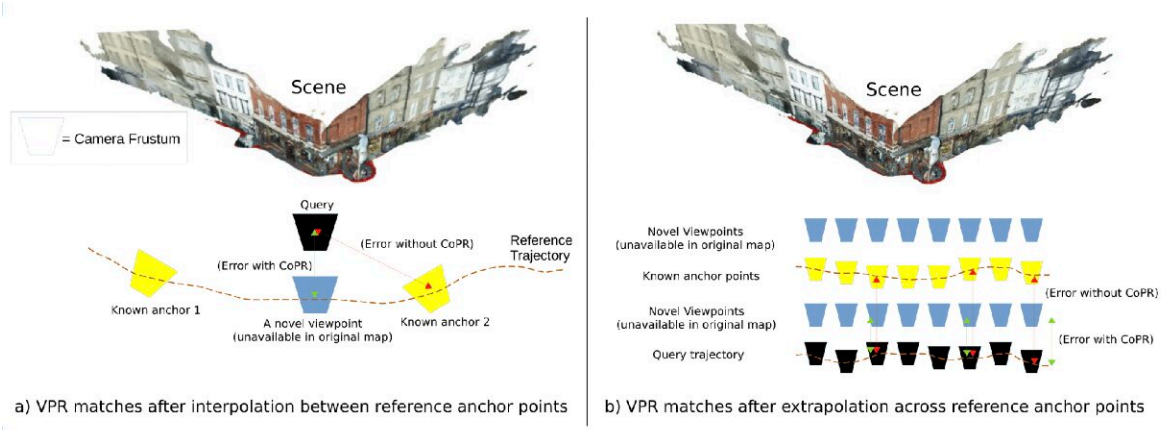

Figure 1: The discrete treatment of Visual Place Recognition (VPR) that leads to lower localization accuracy. Provided that only the yellow anchor reference poses are available in the map, the black query images could only be matched as close as possible to the base error. Regressing descriptors for the blue target viewpoints using interpolation or extrapolation given anchor reference descriptors could lead to improved localization accuracy for query images in VPR and thus reduce the base error.

Credit: Zaffar et al

Visual place recognition (VPR) is the task of identifying the location where specific images were taken. Computer scientists have recently developed various deep learning algorithms that could effectively tackle this task, letting users know where within a known environment an image was captured.

A team of researchers at Delft University of Technology (TU Delft) recently introduced a new approach to enhance the performance of deep learning algorithms for VPR applications. Their proposed method, outlined in a paper in *IEEE Transactions on Robotics*, is based on a new model dubbed continuous place-descriptor regression (CoPR).

"Our study originated from a reflection on the fundamental bottlenecks in VPR performance, and on the related visual localization approaches," Mubariz Zaffar, first author of the study, told Tech Xplore.

"First, we were talking about the problem of 'perceptual aliasing,' i.e., distinct areas with similar visual appearances. As a simple example, imagine we collect reference images with a vehicle driving on the rightmost lane of a highway. If we later drive on the leftmost lane of the same highway, the most accurate VPR estimate would be to match these nearby reference images. However, the visual content might incorrectly match a different highway section where reference images were also collected on the leftmost lane."

One possible way to overcome this limitation of VPR approaches identified by Zaffar and his colleagues could be to train the so-called image descriptor extractor (i.e., a component of VPR models that extracts descriptive elements from images) to analyze images similarly irrespective of the driving lane in which they are taken in. However, this would reduce their ability to effectively determine the place where an image was taken.

"We thus wondered: is VPR only possible if we collect images on all lanes for each mapped highway or if we only drive in the exact same lane? We wanted to extend VPR's simple but effective image retrieval paradigm to handle such practical problems," Zaffar said.

"Second, we realized that even the pose estimate of a perfect VPR

system would be limited in accuracy, as the finite size of the reference images and their poses meant that the map cannot contain a reference with the exact same pose for every possible query, We therefore considered that it might be more important to address this sparsity, rather than trying to build even better VPR descriptors."

When reviewing previous literature, Zaffar and his colleagues also realized that VPR models are often used as part of a larger system. For instance, visual simultaneous localization and mapping (SLAM) techniques can benefit from VPR approaches to detect so-called loop closures, while coarse-to-fine localization approaches can achieve sub-meter localization accuracy by refining the coarse pose estimates of VPR.

"Compared to these more complex systems, the VPR step scales well to large environments and is easy to implement, but its pose estimate is not that accurate, as it can only return the pose(s) of the previously seen image(s) that best visually match the query," Zaffar said.

"Still, SLAM and relative pose estimation do provide highly accurate pose estimates using the same sparse references images and poses, so how are these approaches fundamentally different from VPR? Our observation is that such techniques built a continuous spatial representation from the references which explicitly relates a pose to the visual features, allowing to reason about the visual content at poses interpolated and extrapolated from the given references."

Based on their observations, the researchers set out to explore whether the same continuous representations attained by SLAM and relative pose estimation approaches could be extended to VPR models operating alone. Conventional VPR approaches work by converting a query image into a single so-called descriptor vector and then comparing it with pre-computed descriptors of while coarse-to-fine localization approaches can

achieve sub-meter localization accuracy by refining the coarse pose estimates of VPR. Collectively, all these reference descriptors are referred to as the "map."

After comparing these descriptors, the model determines which reference descriptor most closely matches the descriptor of the query image. The model thus solves the VPR task by sharing the location and orientation (i.e., pose) of the reference descriptor that is most similar to the descriptor of the query image.

To improve VPR localization, Zaffar and his colleagues simply densify the overall "map" of descriptors by employing deep learning models. Instead of thinking of the descriptors of reference images as a discrete set separate from their poses, their method essentially considers the references as points on an underlying continuous function that relates poses to their descriptors.

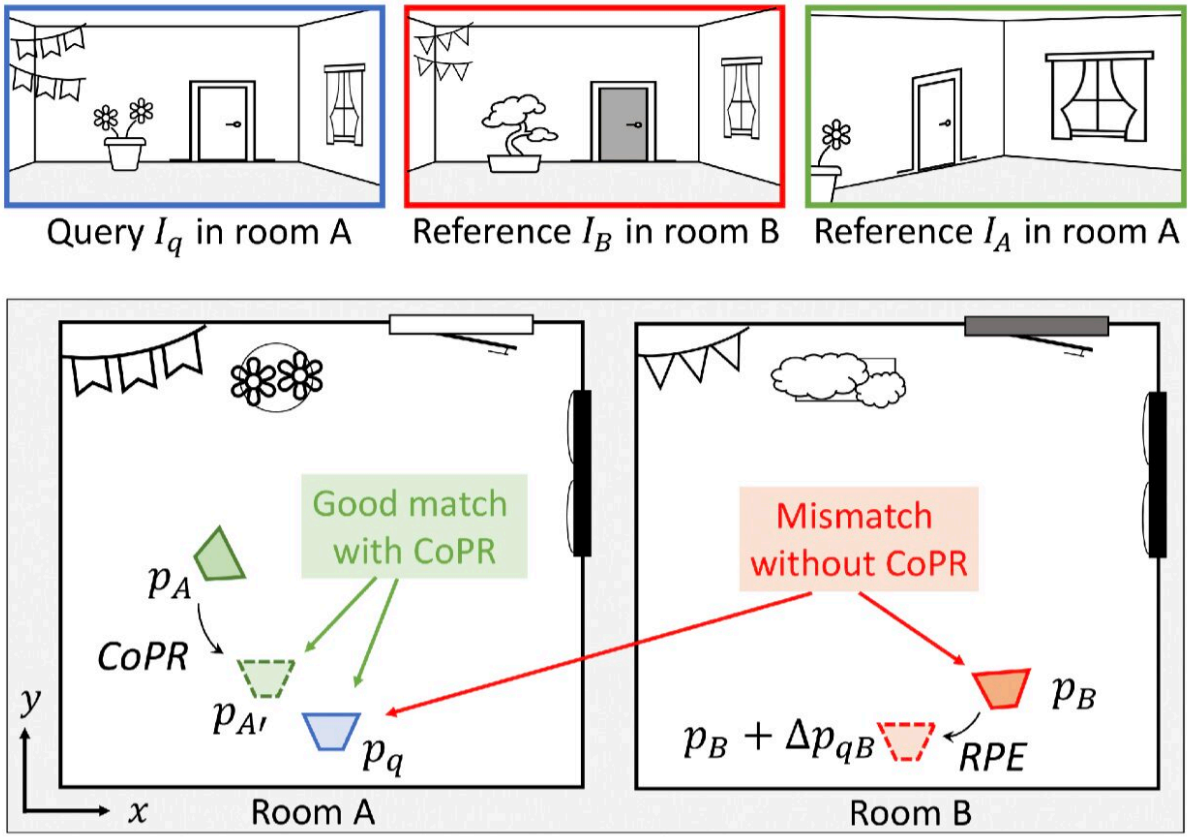

Figure 2: Perceptual aliasing of rooms A and B: query I_q in room A appears more similar to referenc I_B in room B than to reference I_A in correct room A.
If Visual Place Recognition (VPR) retrieves the wrong reference f_B for f_q, Relative Pose Estimation (RPE) between f_B and f_q cannot correct this: the `apparent' difference between the query pose p_q an reference pose p_B is nearly zero. CoPR therefore aims to improve VPR instead by adding references for more diverse poses to the map, e.g., $f_{A'}$ for $p_{A'}$.

Credit: Zaffar et al

"If you think of a pair of references with two nearby poses (so, images with somewhat different locations and orientations, but still looking at the same scene), you can imagine that the descriptors are somewhat similar as they represent similar visual content," Julian Kooij, co-author of the study, explained.

"Still, they are also somewhat different as they represent different viewpoints. While it would be hard to manually define how the descriptors change exactly, this can be learned from the sparsely available reference descriptors with known poses. This is then the essence of our approach: we can model how image descriptors change as a function of a change in pose and use this to densify the reference map. In an offline stage, we fit an interpolation and extrapolation function that can regress the descriptor at an unseen pose from the nearby known reference descriptors."

After completing these steps, the team could densify the map considered by VPR models by adding the regressed descriptors for new poses, which represent the same scene in the reference images but slightly moved or rotated. Remarkably, the approach devised by Zaffar and his colleagues does not require any design alterations to VPR models and allows them to operate online, as the models are offered a larger set of references that they can match a query image to. A further advantage of this new approach for VPR is that it requires relatively minimal computational power.

"Some other recent works (e.g., neural radiance fields and multi-view stereo) followed a similar thought process, also seeking to densify the map without collecting more reference images," Zaffar said. "These works proposed to implicitly/explicitly construct a textured 3D model of the environment to synthesize reference images at new poses, and then densify the map by extracting the image descriptors of these synthetic reference images. This approach has parallels to the 3D point-clouds estimated by visual SLAM, and which requires careful tuning and expensive optimization. Besides, the resulting VPR descriptor could could be inclusive of appearance conditions (weather, seasons etc.) which are considered irrelevant for VPR, or overly sensitive to accidental reconstruction artifacts."

Compared to previous approaches aimed at improving the performance of VPR models by reconstructing the scene in the image space, Zaffar's approach excludes this intermediate image space, which would increase its computational load and introduce irrelevant details. Essentially, instead of reconstructing these images, the team's approach works directly on the reference descriptors. This makes it far simpler to implement to VPR models on a large-scale.

"In addition, our approach does not need to have access to the reference images themselves, it only needs the reference descriptors and poses," Kooij said. "Interestingly, our experiments show that the descriptor regression approach is most effective if a deep learning based VPR method was trained with a loss that weighs descriptor matches on the pose similarity, as this helps align the descriptor space with the geometry of visual information."

In initial evaluations, the researchers' method achieved very promising results albeit the simplicity of employed models, which means that more complex models could soon achieve better performance. In addition, the method was found to have a very similar objective to that of existing methods for relative pose estimation (i.e., for predicting how scenes transform when looking at them from specific angles).

"Both approaches address different types of VPR errors and are complementary," Kooij said. "Relative Pose Estimation can further reduce the final pose errors from a correctly retrieved reference by VPR, but it cannot fix the pose if VPR has incorrectly retrieved the wrong place with a similar appearance to the true location ('perceptual aliasing'). We show with real-world examples that map densification using our method can help identify or avoid such catastrophic mismatches."

In the future, the new approach developed by this team of researchers

could help to agnostically improve the performance of algorithms for VPR applications, without increasing their computational load. As a result, it could also enhance the overall performance of SLAM or coarse-to-fine-localization systems that rely on these models.

So far, Zaffar and his colleagues have tested their approach using simple regression functions to interpolate and extrapolate descriptors, such as linear interpolation and shallow neural networks, which only considered one or a few nearby reference descriptors. In their next studies, they would like to devise more advanced learning-based interpolation techniques that can consider many more references, as this could improve their approach further.

"For instance, for a query looking down a corridor, a reference further down the corridor could provide more detailed information on what the descriptor should contain than a closer reference looking in the other direction," Kooij added.

"Another goal for our future work will be to provide a pretrained map densification network that can generalize to different poses on various datasets, and that works well with little to no finetuning. In our current experiments, we fit the model from scratch on a training split of each dataset separately. A unified pretrained model can use more training data, allowing for more complex network architectures, and give better out-of-the-box results to end-users of VPR."

More information: Mubariz Zaffar et al, CoPR: Toward Accurate Visual Localization With Continuous Place-Descriptor Regression, *IEEE Transactions on Robotics* (2023). DOI: 10.1109/TRO.2023.3262106

© 2023 Science X Network

Citation: A new approach for map densification in visual place recognition (2023, May 22) retrieved 10 May 2024 from https://techxplore.com/news/2023-05-approach-densification-visual-recognition.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.