POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Visual localization in presence of match scarcity

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Valentina Sgarbossa

Advisor: Prof. Luca Magri

Co-advisors: Prof. Giacomo Boracchi, Dr. Antonino Maria Rizzo

Academic year: 2021-2022

1. Introduction

Visual localization consists in inferring the position and orientation from which a given picture (the *query*) was shot, by comparing it to a 3D reconstruction of a scene. As autonomous driving and augmented reality grow their popularity, more and more precise localization is desirable.
Typically, visual localization is performed in two steps: (*i*) matching descriptors of the query and the 3D points of the scene, (*ii*) using the attained correspondences to solve for a pose. To prevent outliers from disrupting the estimate, a Perspective-n-Pose (PnP) solver is used within a robust fitting framework (e.g. RANSAC [2]).
The problem becomes harder when radical appearance variations occur between reconstruction and query (*long-term variations*). State-of-the-art methods [3, 5] employ semantic filters to filter poor matches, because semantics are quite independent of the visual appearance.
Yet, two settings remain critical to these methods: *large-scale* localization, where repetitive structures appear in different locations across the database, and localization with *match scarcity* in long-term scenarios, which entails few correct matches exist overall. In both cases performance drops, respectively by 7.4% and 44.3% of correct localizations.

In this thesis, our contribution is two-fold. First we show these performance drops have a common root cause, i.e. the ambiguity in the space of descriptors impairing matching. Thus, we incorporate semantics within the matching process rather than using it as a post-matching filter. We propose *Semantic Matching*, a matching procedure that considers several candidate matches for all query keypoints, and picks the match with highest semantic consistency. The impact of this shift of perspective is evident in contexts of match scarcity, where filtering – hence throwing away – matches is more harmful than effective.
We furthermore observe that localizing with match scarcity entails worse pose estimates, due to increased chances that random samples achieve larger consensus than the correct pose. The second contribution of our work aims at solving this issue. We adapt the framework of robust pose estimation adding *Biased Consensus*, consisting in evaluating a pose based on the overall semantic consistency of its inliers, rather than their absolute number. We verify the impact of this method on sequences with match scarcity, where Biased Consensus increases performances of over 10% on the baseline.
Finally, combining the proposed tools, we show increased localization ability of over 14% on state-of-the-art algorithms.

2. Problem Formulation

The inputs to visual localization algorithms are a query image and a 3D representation of the scene. We model the information of a query image $\mathrm{I}_q \in \mathbb{R}^{H\times W\times 3}$ with a set of keypoint-descriptor pairs $\mathcal{Q} = \{(\boldsymbol{x}, \boldsymbol{f}) \mid \boldsymbol{x} \in \mathbb{R}^2, \boldsymbol{f} \in \mathbb{R}^D\}$. We also suppose to have a semantic mask $\mathrm{M} \in \{1, ..., L\}^{H\times W}$, with quite a large number of classes L ($\approx 10^2$), hence giving a fine-grained segmentation which is unlikely to overlap significantly to semantic masks of unrelated images.
A reconstruction of the scene, previously obtained from a collection of images $\{\mathrm{I}_d\}$ with Structure from Motion (SfM) can be represented as $\mathcal{X} = \{(\boldsymbol{X_j}, \mathcal{P}_j, \mathcal{V}_j, c_j)\}$, where $\boldsymbol{X_j} \in \mathbb{R}^3$ are the point locations, $\mathcal{P}_j = \{\boldsymbol{d}_1, ..., \boldsymbol{d}_{n_j}\} \subset \mathbb{R}^D$ are two or more associated descriptors, $\mathcal{V}_j$ includes directions and distances of observation of the point at reconstruction time, and $c_j \in \{1, ..., L\}$ is the semantic content.
We assume the camera internal parameters are known for all images. Hence, the goal of localization is to output a pair $(\mathrm{R}, \mathbf{t})$, with $\mathrm{R} \in SO(3)$ representing the camera rotation and $\mathbf{t} \in \mathbb{R}^3$ being the translation vector of the camera centre. The goodness of the pose is measured through the *position error* X_q [m] and *orientation error* Y_q [°] from the available ground-truth pose $(\mathrm{R}_{gt}, \mathbf{t}_{gt})$. These errors are defined as:

$$\mathrm{X}_q = \|\mathbf{C} - \mathbf{C}_{gt}\|_2, \quad \mathbf{C} = -\mathrm{R}^T\mathbf{t} \tag{1}$$

$$\mathrm{Y}_q = |\alpha|, \quad cos\,\alpha = \frac{1}{2}(\mathrm{tr}(\mathrm{R}_{gt}^{-1}\mathrm{R}) - 1) \tag{2}$$

We may assume we know for every image the gravity direction $\boldsymbol{g}_{cam}$ in camera coordinates and the camera height z_0.
A challenge to long-term localization is represented by poor descriptor distinctiveness. This means a query descriptor $\boldsymbol{f}_i$ could have similar or even larger distance to its corresponding descriptors $\boldsymbol{d}_j \in \mathcal{P}_i$ than to unrelated descriptors $\boldsymbol{d}_j \in \mathcal{P}_h$, $h \neq i$.
Moreover, robust pose estimation rests on the assumption that $|\mathcal{I}_{gt}| \gg |\mathcal{I}_{(\mathrm{R},\mathrm{t})}| \;\forall\; (\mathrm{R}, \mathrm{t}) \neq (\mathrm{R}_{gt}, \mathrm{t}_{gt})$, with $\mathcal{I}$ being the inlier set of a pose. However, in long-term settings few correct matches $\{(\boldsymbol{x}_i, \boldsymbol{X}_i)\}$ could be available (*match scarcity*), thus the assumption is likely to be violated.

3. Related Work

We now explore the most relevant research areas for our work, namely descriptor matching techniques and match semantic consistency.

3.1. Matching in Visual Localization

Descriptors of appearance are commonly used as embedding to form 2D-3D correspondences. These descriptors have desirable properties of robustness to scale and rotation changes, and distinctiveness, that is low probability of collision with unrelated descriptors thanks to their high dimensionality.
Matching and filtering are straightforward, since corresponding descriptors are expected to be close to each other, and far from unrelated descriptors. Commonly, all query keypoints are assigned their nearest neighbor in descriptor space. Subsequently, mismatches are filtered out by checking whether the ratio of distances to the first- and second-nearest neighbor is larger than some threshold (*ratio test*). The underpinning principle of this approach is that most correspondences are within two categories: (a) correct matches, and (b) incorrect matches which were formed because the query descriptor has no correspondent among database points.
In large-scale settings a third category should be added, i.e. (c) ambiguous matches. Indeed, the frequent presence of repetitive visual elements entails collisions are more likely to occur due to increasing density in the space of descriptors.
Large-scale localization literature targets these ambiguous descriptors, to improve localization performances. For example, [6] include K-nearest neighbors in the pose estimation, to ensure they pick most correct matches. To keep pose estimation feasible, they let every match vote for some discrete pose and finally select the largest consensus for geometric pose estimation. While this approach works well with repetitive structures, it is not suited to a match scarcity context, where there are too few inliers for the correct model to stand out from background noise. Thus, it is not possible to consider all K neighbors of every match in the pose estimation phase, but we should carefully select at most one match for all keypoints.
Moreover, descriptor ambiguity has broader manifestation than in repetitive structures. It is common in long-term scenarios too, where the

Figure 1: Example of a correct retrieval through Semantic Matching. The considered query keypoint (fuchsia dot) is associated to the wrong location due to the long-term variation of the described patch. The correct match is still close in descriptor space and can be retrieved thanks to the semantic score.

varying appearance pushes matching descriptors away from each other, and closer to unrelated descriptors. Thanks to this observed property, part of the lost matches of long-term localization can be recovered.

3.2. Semantic Match Consistency

Semantic consistency is a notable use of semantics in visual localization. It consists in checking if the labels of 3D points and their 2D correspondents are the same, either locally through the direct comparison of the labels in a 2D-3D correspondence, or globally, through the count of *semantic inliers*, namely the number of projected 3D points that match the label of the pixel they fall into in the query image. These two approaches are known as Simple Semantic Match Consistency [3] (SSMC), and Geometric Semantic Match Consistency [5] (GSMC).
Although providing orthogonal information to appearance, these consistency checks are hardly distinctive, because they rely on coarse segmentations, which may overlap significantly even in unrelated images with homogeneous content (e.g. vegetation). An improved version of the scores [3] employs ad-hoc fine-grained segmentations to reduce the probability of a wrong match to project points to the correct classes.
Both the resulting SSMC and GSMC scores with fine-grained segmentations are used to filter matches to for pose estimation. While this allows to accelerate the research of the correct model by improving the quality of sampled matches, it also reduces the actual amount of matches that are used for pose estimation. In situations of match scarcity this effect is detrimental, as it aggravates the scarcity of correspondences.

4. Proposed method

We present here the two main contributions of this work, *Semantic Matching* and *Biased Consensus*.

4.1. Semantic Matching

We first consider the problem of forming 2D-3D matches, that is find a correspondent in $\mathcal{X}$ for every point in $\mathcal{Q}$.
Let $\boldsymbol{f}_i$ be a query descriptor, and $\{\boldsymbol{d}_1, ..., \boldsymbol{d}_K\}$ the K-nearest neighbor descriptors from the database set, ordered by distance. In large-scale and long-term settings, it is often the case that this order does not rank the desired descriptor first, as observed in Fig. 1. Indeed, it can be seen how the seasonal variation of appearance of the highlighted keypoint (in fuchsia) causes ambiguity in the descriptor space, with the nearest neighbor descriptor coming from an unrelated part of the scene.
We propose to re-rank the descriptors based on global information orthogonal to appearance. For every candidate match $\{\boldsymbol{m}_{i,1}, ..., \boldsymbol{m}_{i,K}\}$ we compute a measure of their quality $q(\boldsymbol{m}_i)$, and choose the match with highest quality. The measure $q(\cdot)$ we choose is

$$q(\boldsymbol{m}_i) = \max_{\phi} \frac{|\mathcal{I}_i^s(\phi)|}{|\mathcal{P}_i^s(\phi)|}, \tag{3}$$

that is a semantic consistency score inspired on the GSMC measure used in [5]. We explain it in the following.
To assess the global semantic consistency, 3D points need to be projected onto the image plane through some pose. Thanks to simplifying assumptions of known gravity direction and camera height, [5] show how to lock all but one degree of freedom in the pose with the information

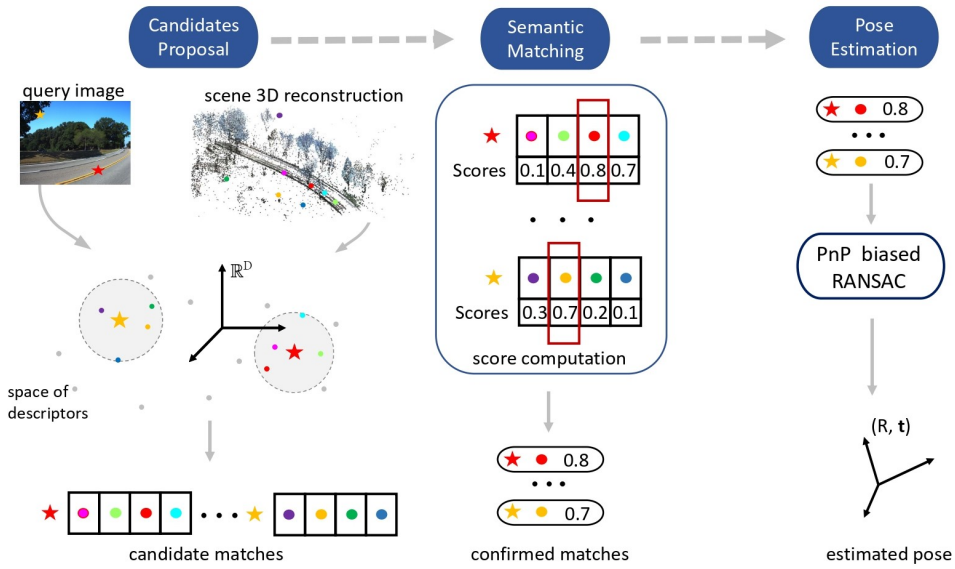


Figure 2: Illustration of Semantic Matching.

of an individual match. The remaining uncertainty is the location of the camera centre on a circle of height z_0, with known orientation. This set of poses can be easily explored parametrizing the position of the circle with discrete angles ϕ. For each angle, one can evaluate the number of semantic inliers $\mathcal{I}_i^s \subset \mathcal{X}$ and the amount of projected points $\mathcal{P}_i^s \subset \mathcal{X}$, and compute the score in Eq. 3.

We choose to evaluate the global semantic consistency as a percentage of inliers over all projected points, differently from [5] who use the absolute inlier count. The choice is motivated by the fact that the number of projected points largely varies from match to match, making the absolute inlier count hardly comparable for different matches.

The final choice of matches is performed by selecting the neighbor $k^* = arg\max_{k=1,\dots,K} q(\boldsymbol{m}_{i,k})$. The set of matches along with their semantic consistency score $\mathcal{M} = \{(\boldsymbol{m}_i, q_i)\}$ is then used directly for biased pose estimation, as in [5].

Fig. 2 provides a summary of the Semantic Matching procedure.

4.2. Biased Consensus for Robust Pose Estimation

Consider a collection of matches $\mathcal{M} = \{(\boldsymbol{m}_i, q_i)\}$ with associated semantic score. To estimate a pose with robust fitting, a RANSAC [2] iteration comprises three phases: (i) sampling correspondences in number equal to the minimal sample size (MSS), (ii) estimation of a hypothesis of pose θ with the PnP method on the sampled points, (iii) evaluation of consensus.

We modify stages (i) and (iii). For (i), we use the semantic scores associated to every match as sampling probabilities, similarly to [5], so to add a bias towards most consistent matches.

We also propose to use semantics in stage (iii) to better assess models in presence of match scarcity. Particularly, we evaluate a model through the total quality of matches that agree with this model. Formally, if $\mathcal{I}_\theta$ are the inliers of a model θ, we evaluate consensus as $CS = \sum_{i \in \mathcal{I}_\theta} q_i$.

The best model will only be updated if it can exhibit matches with better quality than previously sampled poses. Overall, this more conservative estimate is well suited to situations where the best model is not believed to have significantly more consensus than a random sample.

5. Experiments

We conduct two experiments to evaluate both the task of matching and the pose estimation.

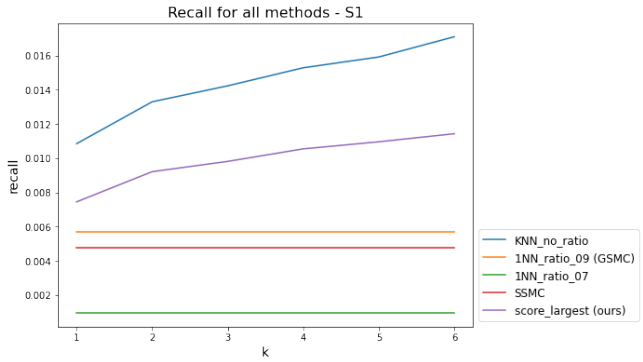
5.1. Dataset and figures of merit

We use the Extended CMU Seasons dataset [1, 4] for the variety of long-term scenarios it contains. All images are taken in Pittsburgh, US. The dataset consists of 11 traversals by car

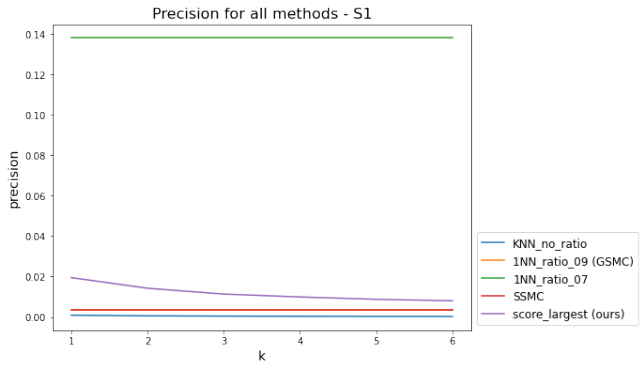
of the same route spanning different seasons and weather conditions, and one reference traversal used for 3D reconstruction, with winter-like appearance and sunny weather. Depending on the dominating content, images are split in Urban, Suburban and Park.
We select three sequences from Park images, with three different levels of complexity: a hard sequence with long-term variations, severe match scarcity and ambiguity of descriptors (S1), a medium sequence without long-term variations but with prevailing vegetation (S2), an easy sequence with abundance of matches and both man-made repeated structures and vegetation (S3). These represent the hardest setting to localize, due to a larger impact of seasonal variations and the repetitiveness of vegetation. We also select a Urban sequence with man-made repeated structures (S4).
We evaluate matching results through precision and recall. Because no ground truth matches are available, we create pseudo-ground truths projecting visible points to image plane with the ground truth poses, and accepting as a valid match every keypoint falling nearer than 5 pixels from the projected point.
Regarding the pose estimation experiments, we count the percentage of queries localized within some threshold of position and orientation error, as defined by Eq. 2. These thresholds are: *fine* 0.25m - 2°, *medium* 0.5m - 5°, *coarse* 5m - 10°.

5.2. Matching experiment

The first experiment evaluates the benefits of increasing the amount of detected matches (recall) by exploring K-nearest neighbors, with respect to false matches (precision). We focus on S1, where match scarcity is mostly impactful.
We compare the following matching strategies: a K-nearest neighbor strategy without any filtering – the maximum we can achieve in terms of recall but with poor precision; 1-nearest neighbor strategies, with ratio test filtering and threshold 0.7 or 0.9 [3, 5], and with SSMC filtering [3]; our matching strategy with top 1 re-ranked match.
The results, shown in Fig. 5.2, indicate that several matches are lost by 1-NN methods, thus confirming the presence of ambiguous keypoints whose matches are found in subsequent neighbors. Our method is well positioned in both precision and recall. The latter is higher than



(a) Recall for considered methods. 1-NN methods are set as constant.



(b) Precision for considered methods. 1-NN methods are set as constant. Note that SSMC and 1-NN ratio 0.9 are overlapping.

other methods even for $k = 1$, which proves the use of filtering strategies is detrimental for the absolute number of retrieved matches.
We additionally search for visual evidence of the increased ability to retrieve correct matches beyond the first. Fig. 1 reports one such example from S1, with the score being the determinant factor allowing to choose the correct neighbor.
The example clearly confirms the validity of the proposed matching strategy, which performs accurate re-ranking of the ambiguous candidate matches.

5.3. Comparison to state of the art

We compare our work to the GSMC and SSMC methods of [3]. The former uses biased pose estimation with semantic consistency scores, and adopts the matching strategy of 1-NN with ratio test at 0.9. The latter filters matches based on simple semantic consistency and estimates a pose with classic RANSAC.
We also include two methods without semantics, to validate the importance of semantics overall.

Method / Setting	Park hard(S1)	Park medium (S2)	Park easy (S3)	Urban (S4)
m	0.25/0.5/5	0.25/0.5/5	0.25/0.5/5	0.25/0.5/5
deg	2/5/10	2/5/10	2/5/10	2/5/10
1-NN + ratio test 0.9 unbiased	0.0/0.0/0.0	30.7/42.0/49.3	32.7/36.4/44.5	70.0/75.5/88.2
k-NN + ratio test 0.9 unbiased	0.0/0.0/0.0	22.0/27.3/37.3	37.3/39.1/47.3	67.3/81.8/**95.5**
SSMC [3]	0.0/0.0/0.0	32.7/43.3/52.0	38.2/48.2/78.2	68.2/85.5/93.6
GSMC [3]	**2.9**/2.9/5.7	46.7/61.3/70.7	47.3/59.1/79.1	71.8/81.8/86.4
sem. matching unbiased	0.0/0.0/0.0	14.0/20.7/25.3	20.0/27.3/39.1	60.0/70.0/83.6
sem. matching biased sampling	**2.9**/4.3/8.6	**50.0**/**63.3**/**72.0**	**60.0**/**73.6**/**92.7**	**81.8**/**89.1**/91.8
sem. matching biased s. + c.	**2.9**/**7.1**/**20.0**	45.3/**63.3**/**72.0**	37.3/54.5/90.0	79.1/87.3/92.7

Table 1: Pose estimation results comparison on all sequences.

These are 1-NN and k-NN with ratio test at 0.9. Regarding our methods, we test the Semantic Matching on pose estimation without bias, thus discarding scores after matching, and with bias as in [5]. On the latter option, we also add the variation of biased sampling and consensus method we propose in Sec. 4.2. Tab. 1 reports the results of pose estimation in all sequences.
Our methods achieve the best performances across most threshold levels of all sequences. The gain is evident in S1, where match scarcity makes it impossible for all unbiased pose estimation techniques to localize correctly. The proposed matching strategy allows to recover correct matches which are decisive for estimating the correct pose. Moreover, the biased sampling and consensus strategy shows outstanding results on this setting, confirming the assumption that the best model is not always the one with largest consensus, in presence of match scarcity.

6. Conclusions

In this work, we have presented novel strategies based on semantics to make long-term visual localization possible in situations of severe match scarcity. The proposed matching strategy, which re-ranks several candidate correspondences of every keypoint through a global measure of the semantic consistency of the matches, is found to improve both the quality and quantity of correct found matches.
Thanks to the improved matches, as well as an enhanced pose fitting algorithm specific to the context of match scarcity, we outperform state-of-the art techniques on selected sequences that exhibit our problem.
Interesting directions for future work include exploiting the pose information associated to semantic scores to accelerate the pose estimation phase, and exploring the potential of visual attention to provide lightweight global cues to perform re-ranking of candidate correspondences.

References

[1] H. Badino, D. Huber, and T. Kanade. The CMU Visual Localization Data Set. `http://3dvis.ri.cmu.edu/data-sets/localization`, 2011.

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981.

[3] M. Larsson, E. Stenborg, C. Toft, L. Hammarstrand, T. Sattler, and F. Kahl. Fine-Grained Segmentation Networks: Self-Supervised Segmentation for Improved Long-Term Visual Localization, 2019. arXiv:1908.06387.

[4] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *CVPR*, 2018.

[5] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl. Semantic match consistency for long-term visual localization. In *ECCV*, 2018.

[6] B. Zeisl, T. Sattler, and M. Pollefeys. Camera pose voting for large-scale image-based localization. In *2015 ICCV*, 2015.