sensors

Article

Appearance-Based Sequential Robot Localization Using a Patchwise Approximation of a Descriptor Manifold

Alberto Jaenal, Francisco-Angel Moreno * and Javier Gonzalez-Jimenez

Machine Perception and Intelligent Robotics Group (MAPIR), Department of System Engineering and Automation Biomedical Research Institute of Malaga (IBIMA), University of Malaga, 29071 Málaga, Spain; ajaenal@uma.es (A.J.); javiergonzalez@uma.es (J.G.-J.)

* Correspondence: famoreno@uma.es

Abstract: This paper addresses appearance-based robot localization in 2D with a sparse, lightweight map of the environment composed of descriptor–pose image pairs. Based on previous research in the field, we assume that image descriptors are samples of a low-dimensional Descriptor Manifold that is locally articulated by the camera pose. We propose a piecewise approximation of the geometry of such Descriptor Manifold through a tessellation of so-called *Patches of Smooth Appearance Change* (PSACs), which defines our *appearance map*. Upon this map, the presented robot localization method applies both a Gaussian Process Particle Filter (GPPF) to perform camera tracking and a Place Recognition (PR) technique for relocalization within the most likely PSACs according to the observed descriptor. A specific Gaussian Process (GP) is trained for each PSAC to regress a Gaussian distribution over the descriptor for any particle pose lying within that PSAC. The evaluation of the observed descriptor in this distribution gives us a likelihood, which is used as the weight for the particle. Besides, we model the impact of appearance variations on image descriptors as a white noise distribution within the GP formulation, ensuring adequate operation under lighting and scene appearance changes with respect to the conditions in which the map was constructed. A series of experiments with both real and synthetic images show that our method outperforms state-of-the-art appearance-based localization methods in terms of robustness and accuracy, with median errors below 0.3 m and 6°.

Keywords: appearance-based localization; computer vision; Gaussian processes; manifold learning; robot vision systems; indoor positioning; image manifold; descriptor manifold

check for updates

Citation: Jaenal, A.; Moreno, F.-A.; Gonzalez-Jimenez, J. Appearance-Based Sequential Robot Localization Using a Patchwise Approximation of a Descriptor Manifold. *Sensors* **2021**, *21*, 2483. https://doi.org/10.3390/s21072483

Academic Editor: Simon Tomažič

Received: 4 March 2021
Accepted: 29 March 2021
Published: 2 April 2021

1. Introduction

Visual-based localization involves estimating the pose of a robot from a query image, taken with an on-board camera, within a previously mapped environment. The widely adopted approach relies on detecting local image features (e.g., points, segments) [1,2] that are projections of 3D physical landmarks. Though this feature-based localization has achieved great accuracy in the last years [3–5], it presents two major drawbacks that hinder long-term localization and mapping: (i) lack of robustness against image radiometric alterations; (ii) inefficiency of 2D-to-3D matching against large-scale 3D models [6]. A much less explored alternative to feature-based localization consists in localizing the robot through the scene appearance, represented by a descriptor of the whole image. According to this framework, localization is accomplished by comparing the appearance descriptor against a map composed of descriptor–pose pairs, without any explicit model of the scene's geometric entities [7,8]. This approach turns out to be particularly robust against perceptual changes and also appropriate for large-scale localization, as demonstrated by the fact that it is included in the front-end of state-of-the-art Simultaneous Localization and Mapping (SLAM) pipelines to perform relocalization and loop closure, typically in the form of Place Recognition (PR) [3,5].

The accuracy of appearance-based localization is, however, quite limited. Good results are reported only when the camera follows a previously mapped trajectory (i.e., in one dimension) [9–11] or when it is very close to any of the poses of the map [8,12]. In this work, we investigate whether localization based only on appearance can deliver *continuous* solutions that are accurate and robust enough to become a practical alternative to methods based on 3D geometric features. We restrict this work to planar motion, that is, we aim to estimate the camera pose given by its 2D position and orientation (3 d.o.f.).

To this end, we first assume that images acquired in a certain environment are samples of a low-dimensional Image Manifold (IM) that can be locally parameterized (or articulated) by the camera pose. This assumption has been justified by previous works [13,14], but only exploited under unrealistic conditions, where the IM was sampled from a fine grid of poses in the environment under fixed lighting conditions. This IM is embedded in an extremely high-dimensional space: $\mathbb{R}^{H \times W \times 3}$, where W and H stand for the width and height of the images, respectively. Patently, working in such extremely high-dimensional space is not only unfeasible, but also impractical since it lacks radiometric invariance. That is, the IM of a given environment might change drastically with the scene illumination and the automatic camera accommodation to light (e.g., gain and exposure time). Thus, it is primarily to project the images (i.e., samples of the IM) to a lower-dimensional space with a transformation that also provides such radiometric invariances [15]. This projection can be carried out by encoding the image into a descriptor vector, hence obtaining a new appearance space, the Descriptor Manifold (DM), which is still articulated by the camera pose. In this paper, we leverage Deep Learning (DL)-based holistic descriptors [8,16] to project the IM into a locally smooth DM. We are aware that such smoothness is not guaranteed for the descriptors employed here, since this feature has not been explicitly taken into account in their design. This issue will be addressed in future work, but in the context of our proposal, the selected DL-based descriptors perform reasonably well under this assumption.

Another capital aspect in appearance-based localization is that it requires an appropriate map, which, in our case, is built from samples of the DM that are annotated with their poses. In this paper, we assume that such samples, in the form of descriptor–pose image pairs, are given in advance and are representative of the visual appearance of the environment. Upon this set of pairs, we propose creating *Patches of Smooth Appearance Change* (PSACs), that is, regions that locally approximate the geometry of the DM using neighbor samples (see Figure 1). A tessellation of such PSACs results in a piecewise approximation of the DM that constitutes our *appearance map*, where pose data is only available at the vertices of the PSACs. The appearance smoothness within each PSAC allows us to accurately regress a descriptor for any pose within the pose space covered by the PSAC. This is accomplished through a Gaussian Process (GP), which delivers the Gaussian distribution of the regressed descriptor (refer to Figure 1).

Our proposal solves continuous sequential localization indoors by tracking the robot pose using a Gaussian Process Particle Filter (GPPF) [17,18] within the described *appearance map*. The particles are propagated with the robot odometry and weighted through the abovementioned Gaussian Process, which is implemented as the GPPF observation model for the image descriptor.

Pursuing to improve the robustness of our method against appearance changes, we model the descriptor variations in such situations as a white noise distribution that is introduced into the estimation of the observation likelihood. Finally, it is worth mentioning that our proposal can easily recover from the habitual PF particle degeneracy problem by launching a fast and multihypothesis camera relocalization procedure through global Place Recognition.

Our localization system has been validated with different indoor datasets affected by significant appearance changes, yielding notable results that outperform current state-of-the-art techniques, hence demonstrating its capability to reduce the gap between feature-based and appearance-based localization in terms of accuracy, while still leveraging the invariant nature of holistic descriptors.

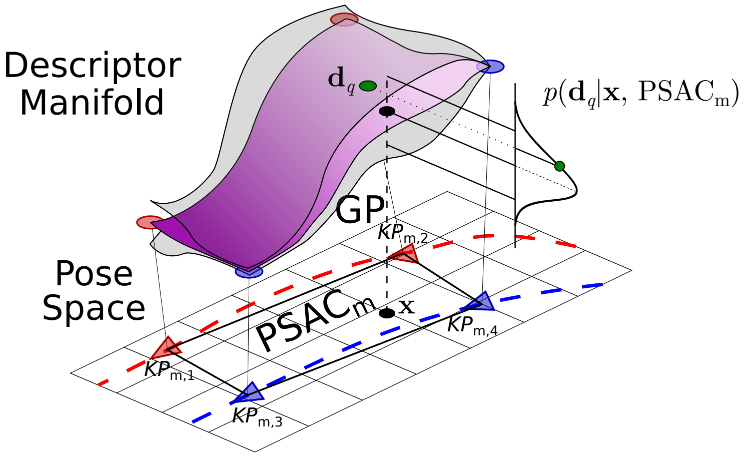


Figure 1. The Gaussian Process (GP) associated to a *Patch of Smooth Appearance Change* (PSACs) approximates the geometry of a neighborhood of the Descriptor Manifold (assumed to be locally smooth) with respect to the pose space, predicting the local likelihood $p(\mathbf{d}_q|\mathbf{x}, \mathrm{PSAC}_m)$ of the observation $\mathbf{d}_q$ in a given pose $\mathbf{x}$. In this example, the descriptor–pose pairs are extracted from two previous trajectories of the robot (in red and blue).

2. Related Work

This section reviews two concepts that are essential for the scope of this work: Global image descriptors and Appearance-based localization.

2.1. Global Image Descriptors

A well-founded way of getting a consistent dimensionality reduction from the Image Manifold to the Descriptor Manifold is through *Manifold Learning* tools, like LLE [19] or Isomap [20]. Their performance, however, is limited to relatively simple IMs that result from sequences of quasi-planar motions or deformations, like face poses, person gait, or hand-written characters [21]. Unfortunately, images taken in a real 3D scene give rise to complex, highly twisted IMs, which also present discontinuities due to occlusion borders [22]. Moreover, typical *Manifold Learning* tools are hardly able to generalize their learned representations to images captured under different appearance settings [15]. This prevents their application to generating low-dimensional embeddings adequate for camera localization.

Nevertheless, Deep Learning (DL)-based holistic descriptors have recently proven their suitability to enclose information from complete images, effectively reducing their dimensionality, while adding invariance to extreme radiometric changes [8,12,23,24]. This feature has made DL-based descriptors highly suitable for diverse long-term robot applications [9,25], e.g., Place Recognition (PR), where the goal is to determine if a certain place has been already seen by comparing a query image against a lightweight set of images [26]. In addition, these descriptors have proven to more sensitively reflect pose fluctuations than local features [26], which is described, for example, by the *equivariance* property [27,28]. Since we are targeting robust robot operation under different appearance conditions, global descriptors arise as the natural choice to address appearance-based localization.

2.2. Appearance-Based Localization

Appearance-based localization is typically formulated as a two-step estimation problem: first, PR is performed to find the most similar images within the map and, subsequently, the pose of the query image is approximated from the location of the retrieved ones [29,30]. In this scenario, DL-based works have proposed to improve the second stage through Convolutional Neural Network architectures that estimate relative pose transformations between covisible images [31–33].

The addition of temporal and spatial sequential information to appearance-based localization methods based on single instances provides more consistency to the estimation of the pose, as it reduces the possibility of losing camera tracking due to, for instance, perceptual aliasing [26]. Following this idea, SeqSLAM [10,34] proposes a sequence-to-sequence matching framework that reformulates PR with the aim to incorporate sequentiality, leading to substantial improvements under extreme appearance changes. Building upon SeqSLAM, SMART [11] integrates odometry readings to provide more consistent results. More recently, *Network-Flow*-based formulations have also been proposed to solve appearance-based sequence matching under challenging conditions, addressing camera localization [35,36] or position-based navigation and mapping [37]. Despite their relevant results, the nature of all these works is discrete, unlike our proposal, restricting all possible estimation to the locations present on the map.

Conversely, CAT-SLAM [38] employs image sequentiality as a source of topometric information to improve the discrete maps used by FAB-MAP [9], allowing interpolation within the sequence map through the association of continuous increments on appearance and pose. Although the estimates produced by this approach are continuous, they are restricted to the mapped trajectory. Our work overcomes this constraint by requiring multiple sequences or pose grids as a source for constructing the map PSACs. This way, we can perform localization even at unvisited map locations near the PSACs, achieving, consequently, more accuracy and reliability.

An interesting alternative to pose interpolation is the use of Gaussian Processes (GPs) regression [39], nonparametric, general-purpose tools that allow generalizing discrete representations to a continuous model, and, hence, can be adapted to perform continuous localization within discrete maps. For instance, [40,41] employed GPs to generate position estimates for omnidirectional images in indoor maps, achieving good performance although lacking applicability to robot rotations. Our approach is instead designed to work with both 2D positions and rotations for conventional cameras.

In turn, the authors of [18] proposed Gaussian Process Particle Filters (GPPFs) to solve appearance-based localization in maps of descriptor–pose pairs. The GP works as the observation model of the PF, estimating the likelihood of the observed holistic descriptor at each of the particle positions. This localization pipeline was later improved in [42] by using only the nearest map neighbors in the GP regression, allowing efficient localization within large environments. Despite being promising, both works have three major drawbacks: (i) they define a unique Gaussian Process between poses and descriptors for the whole environment, assuming that the manifold geometry has a similar shape for the entire environment, thus leading to inaccurate estimations, (ii) they do not propose a relocalization process in case of losing tracking, and (iii) they only consider localization under the same appearance than the map, lacking robustness to radiometric alterations.

Inspired by these works, we employ a GPPF to solve appearance-based localization in a continuous and sequential fashion within challenging indoor environments. We solve their first problem by locally modeling the mapping between poses and descriptors via specific GPs for each PSAC, providing refined estimates for each neighborhood. Our proposal solves the second issue through a fast and multihypothesis relocalization process based on global PR within the map. Finally, the last issue is addressed by incorporating a model of the appearance variation between the mapped and query images to the map.

3. System Description

This section describes our proposal for the process of appearance-based camera localization. First, we define the elements that form the *appearance map*, which are key contributions in this work, and then we address PR-based localization and camera tracking using a probabilistic formulation based on a GPPF.

3.1. Patches of Smooth Appearance Change

The *Patches of Smooth Appearance Change* (PSACs) are regions that locally model the interrelation between camera poses and image descriptors, and represent areas where the change in appearance is small.

3.1.1. Definition

The basic building unit of a PSAC is the pair $p_i = (\mathbf{d}_i, \mathbf{x}_i)$, composed by the global descriptor $\mathbf{d}_i \in \mathbb{R}^d$ of an image and the pose $\mathbf{x}_i \in SE(2)$ where it was captured.

We assume that these pairs are extracted from any of these two environmental representations (Figure 2): either from several **robot navigation sequences** (at least two) or from **pose grids** where the cameras have densely sampled the environment given fixed position and rotation increments (i.e., a regular grid). Optimally, a subset of these pairs should be selected so that they constitute the smallest number of samples from which the Descriptor Manifold (DM) can be approximated with sufficiently good accuracy. These *key* samples can be viewed as the equivalent to the *key-frames* in traditional, feature-based visual localization, and hence, we denote them *key-pairs* (*KPs*). Since determining such an optimal subset is a challenging issue itself, out of the scope of this work, the *KPs* are constantly sampled from the total collection of pairs.

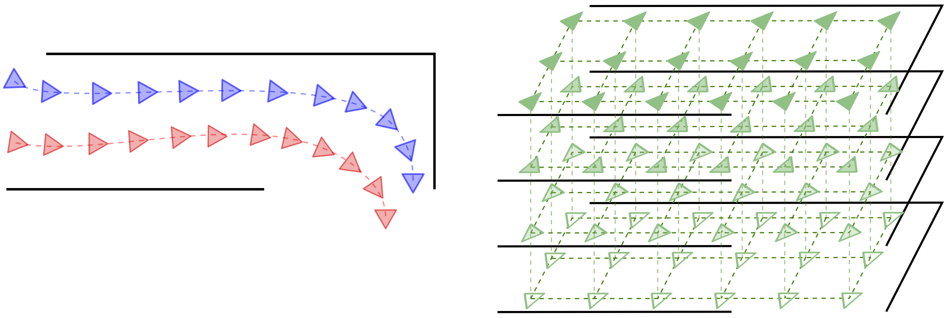


Figure 2. Each PSAC is constructed from descriptor–pose pairs that can be obtained from two different robot trajectories (left, blue and red) or a grid of poses (right, green). This is better seen in color.

Each PSAC is built from a group of adjacent *KPs* and approximates the DM in the region that they delimit. As explained later, the robot localization takes place within these PSACs by defining a suitable observation model for the GPPF (refer to Figure 1). Formally, let the m-th PSAC be

$$\text{PSAC}_m = \left(\left\{KP_{m,i} | \, i = 1, \ldots, Q\right\}, GP_m\right), \tag{1}$$

where $Q \geq 3$ is the number of *KPs* forming the PSAC. In turn, GP_m is a Gaussian Process specifically optimized for that particular PSAC that delivers a Gaussian distribution over the image descriptor for any pose nearby the PSAC (further explained in Section 3.1.2).

Thus, in order to determine the closeness between a query pair $p_q = (\mathbf{d}_q, \mathbf{x}_q)$ and a particular PSAC, we define two distance metrics as follows:

- The *appearance distance* $D^a_{m,q}$ from the query descriptor $\mathbf{d}_q$ to the m-th PSAC is defined as the average of the descriptor distances to each of its constituent *key-pairs*:

$$D^a_{m,q} = D^a(\text{PSAC}_m, \mathbf{d}_q) = \frac{1}{Q}\sum_i^Q ||\mathbf{d}_q - \mathbf{d}_{m,i}||_2. \tag{2}$$

- Similarly, but in the pose space, we define the *translational distance* $D^t_{m,q}$ from $\mathbf{x}_q$ to the m-th PSAC as

$$D^t_{m,q} = D^t(\text{PSAC}_m, \mathbf{x}_q) = \frac{1}{Q}\sum_i^Q ||\mathbf{t}_q - \mathbf{t}_{m,i}||_2, \tag{3}$$

with $\mathbf{t}_q$ being the translational component of the pose $\mathbf{x}_q = (\mathbf{t}_q, \theta_q)$.

Finally, the set of all PSACs covering the environment that has been sampled forms the **appearance map** $\mathcal{M}$:

$$\mathcal{M} = \{\text{PSAC}_m |\, m = 1, \dots, M\}, \tag{4}$$

with M being the number of PSACs. This way, we achieve a much more accurate approximation of the relation between the pose space and the DM within each PSAC, ultimately modeling in $\mathcal{M}$, patchwise, the complete shape of the DM.

3.1.2. GP Regression

GPs are powerful regression tools [39] that have previously demonstrated their validity as observation models in Particle Filters [17,18,42].

In this work, we learn a specific GP for each PSAC from its vertices *KPs* and also using all the nearest pairs, in terms of translational distance. Then, for a certain query pose $\mathbf{x}_q$, the GP_m delivers an isotropic Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{m,q}, \sigma^2_{m,q}\mathbf{I}_d)$, where $\boldsymbol{\mu}_{m,q} \in \mathbb{R}^d$ and $\sigma^2{}_{m,q} \in \mathbb{R}$ stand for its mean and uncertainty, respectively. This distribution is finally employed to estimate the likelihood $p(\mathbf{d}_q|\mathbf{x}_q, \text{PSAC}_m)$ of an observed image descriptor $\mathbf{d}_q$, given the query pose $\mathbf{x}_q$ within the PSAC$_m$.

For this, the GP regression employs a *kernel* k, which measures the similarity between two input 2-D poses $(\mathbf{x}_i, \mathbf{x}_j)$, with this structure:

$$k(\mathbf{x}_i, \mathbf{x}_j) = k_{RBF}(\mathbf{t}_i, \mathbf{t}_j) \cdot k_{RBF}(\theta_i, \theta_j) + k_W(\mathbf{x}_i, \mathbf{x}_j). \tag{5}$$

This *kernel* k first multiplies two Radial Basis Function (RBF) *kernels* $k_{RBF}(a_i, a_j) = \beta_a \exp(-\alpha_a ||a_i - a_j||^2_2)$ (α_a and β_a are optimizable parameters) for the separated translational and rotational components of the evaluated poses $\mathbf{x} = (\mathbf{t}, \theta)$. Then, a White Noise *kernel* $k_W(a_i, a_j) = \sigma^2_W \delta(a_i - a_j)$ is added, which models the variation suffered by the image descriptors taken at the same pose but under different appearances (refer to Figure 3). This is justified because, although global PR descriptors have demonstrated outstanding results in terms of invariance, such invariance is not ideal and small differences might appear. Thereby, since the construction of the map $\mathcal{M}$ is typically carried out considering just one particular appearance, and we aim for the robot localization to be operational under diverse radiometric settings, we propose the inclusion of a white noise distribution accounting for this circumstance in the regression. We model such descriptor variation with the variance σ^2_W, computed as the average discrepancy between the descriptor variances of pose adjacent pairs p_i, under the same $\sigma^2_{i,\text{same}}$ and different $\sigma^2_{i,\text{diff}}$ illumination settings:

$$\sigma^2_W = \frac{1}{N}\sum_i^N \left(\sigma^2_{i,\text{diff}} - \sigma^2_{i,\text{same}}\right). \tag{6}$$

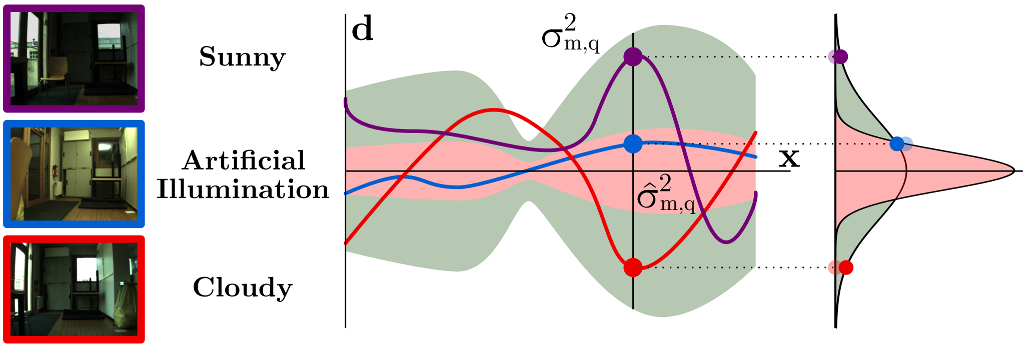


Figure 3. These three images have been captured at the same pose but with different appearances. Ideally, their descriptors (blue, red, and purple dots) should be identical but, in practice, certain inaccuracies appear. As the GP learns the descriptor distribution uniquely from the appearance of the map, this variation is not considered, leading to an underestimated GP uncertainty (red area, $\hat{\sigma}^2_{m,q}$). The inclusion of white noise expands such uncertainty (green area, $\sigma^2_{m,q}$), solving this issue.

3.2. Robot Localization

Once we have defined all the elements involved in the representation of the environment, we address here the process of localization within the appearance map $\mathcal{M}$. We aim to estimate the robot pose through appearance-based continuous tracking using a Gaussian Process Particle Filter (GPPF), namely, a PF that employs the GPs described above as observation models. Being well-known robot localization tools, we do not provide a deep description of Particle Filters here but instead refer to the reader to the seminar work [43] for further information.

At time-step t, each of the P particles in the filter represent a robot pose estimation $\mathbf{x}_p^{(t)}$ with an associated weight $w_p^{(t)}$, proportional to its likelihood. Besides, each particle is assigned to a certain region $\text{PSAC}_{m,p}^{(t)}$, as explained next.

3.2.1. System Initialization

When the PF starts, we perform global localization based on Place Recognition to select the most similar PSAC in $\mathcal{M}$ to the query descriptor according to its appearance:

$$\text{PSAC}_{\hat{m}} = \min_{\text{PSAC}_m \in \mathcal{M}} D_{m,q}^{a}. \tag{7}$$

To account for multihypothesis initialization, we also consider as candidates those PSACs whose *appearance distance* is under a certain threshold proportional to $D_{\hat{m},q}^{a}$. Subsequently, the particles are uniformly assigned and distributed among all candidate PSACs, setting their initial weights to $w_p^{(t_0)} = \frac{1}{P}$.

Note that, if the robot tracking is lost during navigation, this procedure is launched again to reinitialize the system and perform relocalization.

3.2.2. Robot Tracking

Once each particle is assigned to a candidate PSAC, the robot pose estimation is carried out following the traditional *propagation-weighting* sequence:

Propagation. First, the particles are propagated according to the robot odometry:

$$\mathbf{x}_p^{(t)} = \mathbf{x}_p^{(t-1)} \oplus v^{(t)}, \tag{8}$$

with $v^{(t)} \sim \mathcal{N}(\overline{v}^{(t)}, \Sigma_v)$ representing noisy odometry readings, and $\oplus$ being the pose composition operator in $SE(2)$ [44].

Weighting. After the propagation, the *translational distance* between each particle's pose and all the PSACs is computed, so that the particle is assigned to the nearest PSAC ($\text{PSAC}_{m,p}^{(t)}$). Then, we use the GP regressed in the PSAC to locally evaluate the likelihood of the observed descriptor $\mathbf{d}_q$ at the particle pose $\mathbf{x}_p^{(t)}$ as follows:

$$w_p^{(t)} = p\big(\mathbf{d}_q | \mathbf{x}_p^{(t)}, \text{PSAC}_{m,p}^{(t)}\big) \sim \exp\Bigg(-\frac{d}{2} \ln(\sigma 2_{m,p}^{(t)}) - \frac{||\mathbf{d}_q - \boldsymbol{\mu}_{m,p}^{(t)}||_2^2}{2(\sigma 2_{m,p}^{(t)})} \Bigg), \tag{9}$$

with d being the dimension of the descriptor.

Finally, apart from *propagation* and *weighting*, two more operations can be occasionally applied to the particles.

Resampling: In order to prevent particle degeneracy, the GPPF resamples when the number of effective particles is too low, promoting particles with higher weights.

Reinitializing: During normal operation, the GPPF may lose the tracking of the camera, mainly due to extremely challenging conditions in the images (e.g., very strong change appearances, presence of several dynamic objects). We identify this situation by inspecting the *translational distance* between each particle and the *centroid* of its assigned

PSAC, defined as the average pose of all the key-pairs forming the PSAC. If all particles are at least twice farther from the centroid of their assigned PSAC than its constituent *key-pairs*, the tracking is considered lost. Consequently, the PF relocalizes by following the aforementioned initialization procedure.

4. Experimental Results

In this section, we present three experiments to evaluate the performance of our appearance-based localization system.

First, we carry out a verification of the regression outcome in Section 4.1, with the aim to experimentally validate the hypothesis of a smooth Descriptor Manifold within the regions covered by each PSAC. In Sections 4.2 and 4.3, we test our proposal with four different state-of-the-art global descriptors in two different datasets, with a combination of setups for the map sampling. This has provided us with an insight about the error incurred by our proposal and has allowed us to determine the best configuration for localization. The second experiment, in Section 4.4, compares the resulting setup with three appearance-based localization alternatives in terms of accuracy and robustness, revealing that our system equals or improves their performance in every scenario.

It is important to highlight that this evaluation does not include comparisons with feature-based localization methods, since they are not compatible with appearance changes in the images used for both mapping and localization, which is one of the main benefits of our proposal.

We employed two different indoor datasets for the experiments:

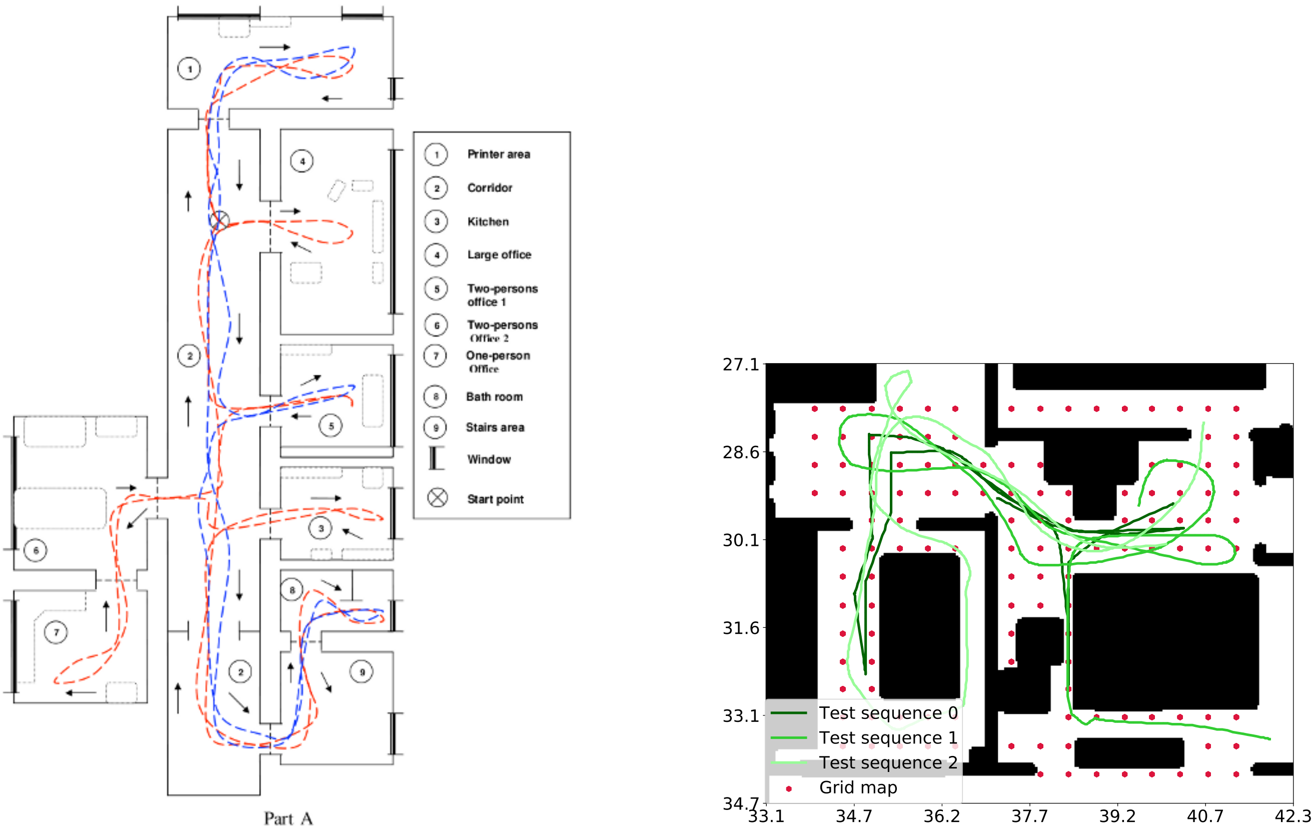
- The **COLD-Freiburg database** [45], which includes real images from an office gathered with a mobile robot under different appearance conditions. We have used only a representative subset of the sequences in *part A* of the dataset (Figure 4a).
- The synthetic **SUNCG Dataset** [46] rendered through the virtual **HOUSE3D** environment [47] allows us to test the localization on a grid map (see Figure 4b).

In turn, regarding the global image representation, we have tested the following state-of-the-art appearance descriptors:

- **NetVLAD**: VGG16 [48] based on off-the-shelf, 4096-sized NetVLAD features with Principal Component Analysis (PCA) whitening [8].
- **ResNet-101 GeM**: ResNet-101-based [49] fine-tuned generalized-mean features with learned whitening [12].
- **1M COLD Quadruplet** and **1M RobotCar Volume**: end-to-end learned condition invariant features with VGG16 NetVLAD [24] with quadruplet and volume loss functions in two different datasets.

Although none of these descriptors has been specifically designed to fulfill suitable properties for our pose regression approach, they have achieved promising results in terms of localization accuracy, as shown next. We used the GPy tool [50] to implement the proposed Gaussian Processes and empirically determined σ_W^2 from Equation (6), for each descriptor, by randomly sampling $N = 2000$ adjacent pairs with diverse illumination settings from the COLD-Freiburg database.

In this evaluation, we employ as metrics the median errors in translation and rotation (to inspect our method's accuracy), as well as the percentage of correctly localized frames (which illustrates the tracking and relocalization capabilities of our method). It is worth mentioning that traditional trajectory-based evaluation metrics as Absolute Trajectory Error (ATE) or Relative Pose Errors (RPE) are not applicable to this approach since our proposal, and appearance-based localization methods in general, yields global pose estimations that are not guaranteed to belong to a trajectory, due to possible tracking losses and relocalization situations.



(**a**) (**b**)

Figure 4. Environments and sequences employed for evaluation. (**a**) Map of the COLD-Freiburg Part A environment. Samples of the standard and extended routes are depicted in blue and red, respectively (image from [45]). (**b**) Map of the house rendered by the SUNCG environment (the house employed was *034e4c22e506f89d668771831080b291*). The dense grid poses are shown in red and the test sequences in different shades of green. Black regions depict objects, where the robot is not able to navigate.

4.1. Corridor: Sanity Check

The main assumption of our proposal is the hypothesis of a locally smooth Descriptor Manifold with respect to the pose, on which PSACs are based. Since this assumption is not justified by previous work, we have conducted a basic test to evaluate the regression outcome of the PSACs in a simple scenario.

The proposed experiment studies the evolution of the image descriptor along a simple, linear trajectory, by comparing the observed descriptor and the mean of the descriptor distribution resulting from the regression within the PSAC. In this manner, the behavior of the descriptor can be examined along the corridor axis in order to prove its continuity and the validity of the PSAC approximation. For this, we have selected a portion of an artificially illuminated (*night*) sequence where the robot traveled along a ∼8 m-long corridor, as well as the NetVLAD image descriptor. For the PSACs, we used a map constructed with images with the same appearance selected every 20 frames.

In order to represent the evolution of the descriptors, we have applied Principal Component Analysis (PCA) to them and represented the first PCA element (that with larger variation). Thus, Figure 5 depicts the trajectory of the robot through the corridor along with the value of said first PCA element for both the observed descriptor and the mean of the descriptor distribution regressed by the GPs at each PSAC. The displayed results demonstrate that the descriptor has a continuous evolution along the corridor, almost lineal in the central part. Besides, the PSACs are proved to also have a continuous outcome and to approximate very accurately the values of the observed descriptor along the sampled trajectory.

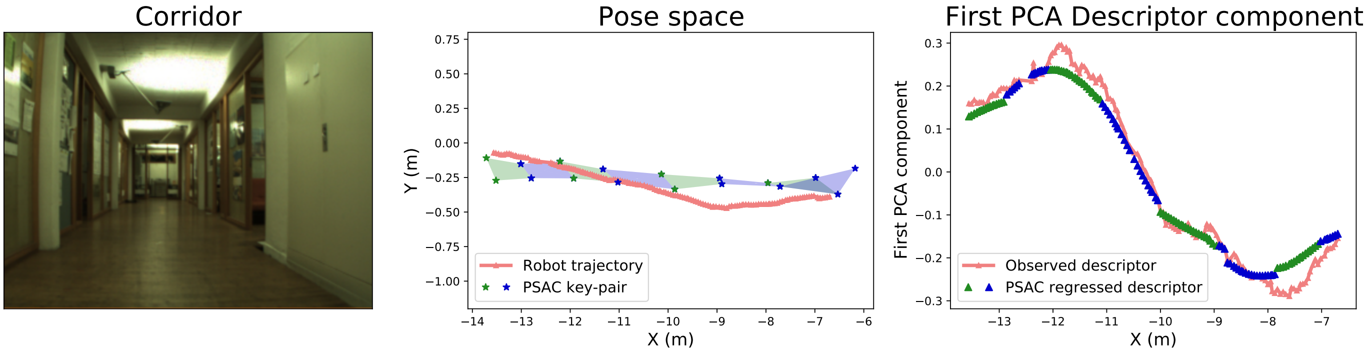


Figure 5. Experimental study of the global image descriptor behavior along a corridor of the COLD-Freiburg database (left image). The central figure depicts the robot pose trajectory and the PSACs (shadowed areas) through which it navigates. The rightmost figure compares, along the corridor, the behavior of the first Principle Component Analysis (PCA) component of the observed descriptor (coral) and the PSAC regression output (blue and green). This is better seen in color.

4.2. COLD: Sequential Map Testing

The COLD-Freiburg database (*part A*) provides odometry readings and real images for two different itineraries, namely, (i) **extended** (∼100 m-long), which covers the whole environment; (ii) **standard** (∼70 m), covering a subset of the environment, both depicted in Figure 4a. The dataset also provides images gathered under three different lighting conditions: at *night* (with artificial illumination), and on *cloudy* and *sunny* days.

In order to create the *appearance maps* for the experiments, we employed the first and second *night* sequences of the extended itinerary, since images captured under artificial illumination do not suffer from severe exposure changes or saturation like under the remaining conditions. From here on, we will refer to these as *map sequences*. Specifically, *key-pairs* from both *map sequences* have been obtained through Constant Sampling (CS) every 10, 20, and 30 pairs, resulting in three different maps with diverse density (described in more detail in Table 1), using $Q = 4$ *KPs* to construct every PSAC.

Table 1. Compared statistics of the created appearance maps. The dimension is calculated for a 2048-sized global descriptor.

Dataset	Area	Map	*Key-Pairs* (*KPs*)	PSACs	Size (Mb)	Construction Time (s)
COLD Database	∼900 m^2	Samp. 10	559	321	2.18	56.68
		Samp. 20	280	159	1.09	32.27
		Samp. 30	187	103	0.73	25.92
SUNCG	∼45 m^2	Dense	1203	679	4.69	140.94
		Sparse pos	451	227	1.76	129.36
		Sparse rot	723	432	2.82	136.95
		Sparse pos-rot	271	149	1.05	107.93

Finally, we have setup an extensive evaluation with six other sequences including different routes and illumination conditions: the first *night*; *cloudy* and *sunny* sequences of the **standard** part; the first *cloudy* and *sunny* sequences; and the third *night* sequence of the **extended** part.

Figure 6 shows a comprehensive test study depicting the localization performance of our proposal after twenty runs for all test sequences at each map, using the median translational (top) and rotational (down) errors as metrics. Note that the number of particles for the PF has been set to $P = 10^3$, as we have empirically found that increasing that number does not improve the accuracy results. The overall performance shows a median error predominantly below 0.3 m and 6°, which denotes promising results given the pure appearance nature of our approach, i.e., with no geometrical feature employed for localization.

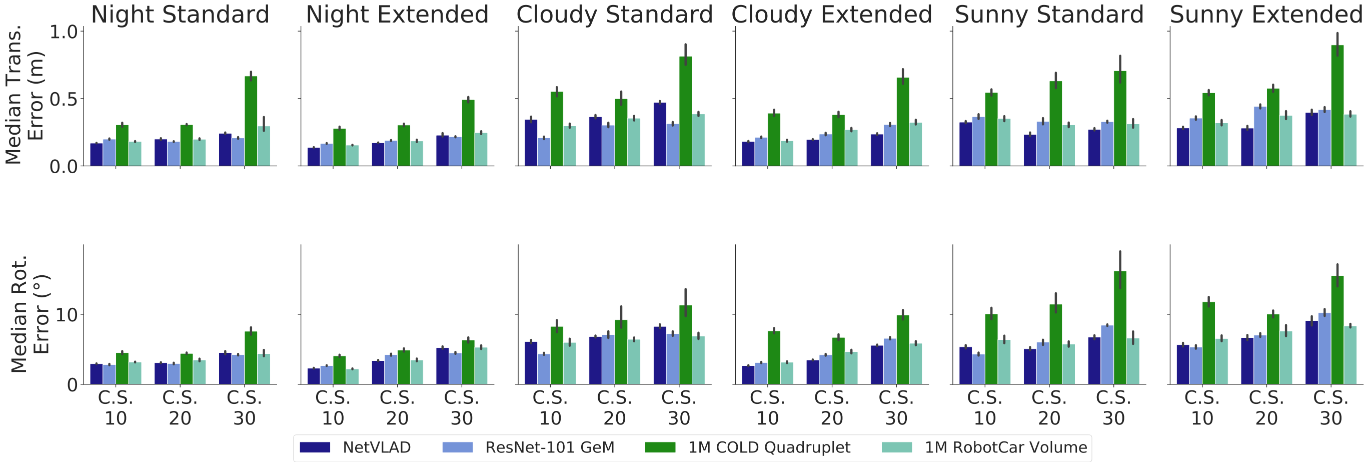


Figure 6. Comparison of the median translational and rotational errors of our proposal in the COLD-Freiburg dataset, tested at every sequence and with different Constant Sampling (CS) rates, using each holistic descriptors with 10^3 particles. Note that the maps were constructed employing sequences under similar conditions to the *night* sequences. This is better seen in color.

The results in Figure 6 show that scene appearance seems to be a key issue regarding the system's accuracy, as our proposal achieves better results in less-demanding lighting conditions like artificial illumination (night) or cloudy. Nevertheless, our system still demonstrates notable performance under challenging radiometric conditions, such as in sunny sequences (e.g., presence of lens flares and image saturation), hence proving its suitability for robust appearance-based localization.

On the other hand, the number of *KPs* that form the map is another factor influencing the performance, since the PSACs approximate the pose–descriptor interrelation the closer their *KPs* are. Although not particularly significant under advantageous conditions, this factor severely affects performance under challenging situations, as in *sunny* sequences, where localization is hindered as the sampling frequency decreases. Note that a more elaborate mapping technique than CS would improve these results, since an optimal selection of *KPs* would conform PSACs that achieve a more precise description of the DM geometry. Nevertheless, this will be explored in future work while, in this paper, we rely on CS to get still nonoptimal but notable results.

Finally, regarding the tested PR descriptors, the results show that in most cases, all perform similarly, with NetVLAD mostly achieving slightly better results. In turn, 1M COLD Quadruplet seems to struggle under complex illumination conditions, which might indicate that its empirically estimated white noise variance is unlikely to account adequately for these cases. The similar performance shown by all descriptors agrees with the fact that none of them has been specifically trained for appearance-based localization.

4.3. SUNCG: Grid Map Testing

The SUNCG Dataset provides a set of synthetic houses where a virtual camera can be placed at any pose. This feature allowed us to create a regular grid map in the space of planar poses with the camera and then to evaluate the impact of the map density in our proposal. Note that this dataset does not present appearance changes, and hence, the effect of such a characteristic cannot be evaluated in this experiment.

First, our *dense* grid map was created by selecting *KPs* with constant increments of 0.5 m in translation and 36° in rotation (refer to the red dots in Figure 4b). Then, we used subsampling to generate more grid maps for the evaluation, as described in Table 1, namely, the *sparse-position map* (subsampling half of the positions); the *sparse-rotation map* (subsampling half of the orientations); and the *sparse-position-rotation map* (subsampling both at the same time). In this case, we created PSACs with $Q = 8$ *KPs* for all maps. Additionally, we have recorded three ∼30 m-long test sequences following the trajectories shown in shades of green in Figure 4b, generating a *synthetic odometry* corrupted by zero-mean Gaussian noise with $\sigma_u = (0.06\,\text{m},\ 1^\circ)$.

The results of this experiment are shown in Figure 7, comparing the median errors in translation (left) and rotation (right) for all the descriptors employed in the previous experiment and for the described versions of the grid map. Again, we have set $P = 10^3$ particles for the PF. As can be seen, our proposal yields median errors under 0.2 m and 6° in the *dense map*, while using subsampled maps hinders the process of localization. It can be noted that subsampling exclusively on rotations does not worsen the accuracy, while subsampling positions has a noticeable impact on the overall performance. Consequently, PSACs prove to handle information sparseness more efficiently in orientation than in position. Not surprisingly, subsampling in both position and orientation clearly achieves the worst localization performance due to the combined loss of information.

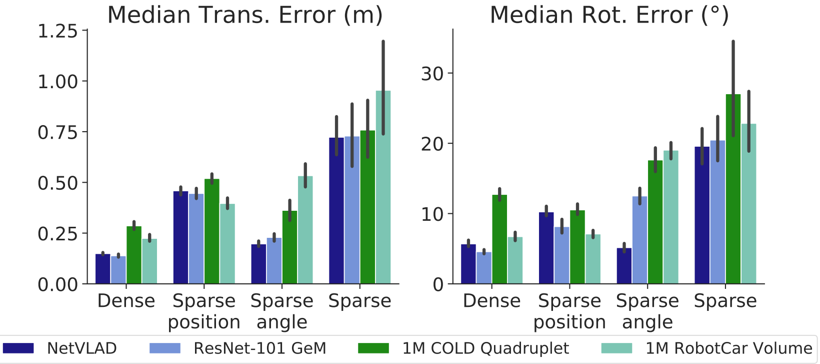


Figure 7. Median tracking errors of our proposal for the test sequences on the SUNCG generated maps, with 10^3 particles.

Regarding the holistic descriptors, all of them again demonstrate a similar behavior for each subsampling case, with 1M RobotCar Volume performing worst. ResNet-101 GeM and NetVLAD, in turn, achieve the best performance.

These results demonstrate that uniform grid sampling is a rough strategy for mapping environments, achieving results highly dependent on the sampling density. Besides, the construction of such maps with real robots becomes largely time-consuming, mostly being realizable when using virtual environments. Future work should investigate more elaborated strategies, designed to fulfill more adequate criteria concerning the map creation, ultimately pursuing an optimal approximation of the DM geometry.

4.4. Comparative Study

Finally, we compare the localization performance between our proposal and state-of-the-art appearance-based methods of diverse nature in both datasets. For the setup of our method, we selected the NetVLAD descriptor due to its performance against appearance changes, and added the *KPs* every 20 pairs for the COLD dataset, as it represents a fair trade-off between accuracy and the number of *KPs* employed.

These are the appearance-based localization methods involved in the comparison:

- **Gaussian Process Particle Filter (GPPF)** [42], configured with $P = 10^3$ particles.
- The **Pairwise Relative Pose estimator (PRP)** presented in [31]: a CNN-based regressor that estimates the pose transform between the query and the 5 most similar map images obtained through PR.
- The **Network flow** solution proposed in [37]: a sequential sparse localization method that includes *uniform* and *flow-based* mapping, both considered in this study. In order to make the results comparable, we modified its outcome, which is sparse, to pro-

duce continuous estimations. For that, we used the following weighting after the bipartite matching:

$$\mathbf{x}_i = \frac{\sum_j^{N_k} \frac{\mathbf{y}_i}{\mathbf{y}_{ij}} \mathbf{x}_i}{\sum_j^{N_k} \frac{\mathbf{y}_i}{\mathbf{y}_{ij}}} \tag{10}$$

where $\mathbf{x}_j$ is the pose of the each of the $N_k = 5$ most contributing *KPs*, $\mathbf{y}_{ij}$ is the flow connecting the i-th query and the j-th *KP*, and $\mathbf{y}_i = \sum_j^{N_k} \mathbf{y}_{ij}$ represents the query flow from the nearer *KPs* (refer to [37] for further info).

- **Our approach**, configured with $P = 10^3$ particles.

Table 2 shows the compared performance between all the described algorithms after twenty runs for each scenario. Note that, apart from the median errors, we included the percentage of correctly localized frames along the trajectory, showing the tracking robustness and relocalization potential. Concretely, a frame is considered to be correctly localized when the distance between the estimate and its true pose is below (0.5 m, 10°).

Table 2. Comparative median position and rotation errors, and % of correctly localized frames (m, $^\circ$, $(\%)$) of different state-of-the-art appearance-based localization methods. A frame is correctly localized when the distance between the estimate and its true pose is below (0.5 m, 10°) (L: sequences where the tracking got lost. N/A: not applicable). PRP—Pairwise Relative Pose estimator; PR—Place Recognition. In bold, best result for each sequence.

Dataset	Map	Sequence	GPPF [18,42] + Unif. Sampl	PRP CNN [31] + NetVLAD PR	Network Flow [37] + Unif. Sampl.	Network Flow [37] + Flow Sampl	Our Method C.S.
COLD Database	Samp. 20	Night std	L	1.17, 10.94 (8%)	**0.19**, 4.33 (66%)	0.26, 4.66 (60%)	0.2, **3.08** (**85**%)
		Cloudy std	L	1.93, 14.59 (4%)	0.31, **4.76** (**57**%)	0.36, 5.29 (46%)	**0.3**, 5.82 (56%)
		Sunny std	L	2.2, 14.75 (3%)	0.36, 6.91 (40%)	0.42, 7.7 (33%)	**0.23**, **5.08** (**66**%)
		Night ext	L	1.27, 11.07 (8%)	0.22, 3.88 (69%)	0.26, 4.36 (61%)	**0.17**, **3.38** (**82**%)
		Cloudy ext	L	1.46, 12.17 (6%)	0.22, 4.13 (60%)	0.31, 5.12 (52%)	**0.2**, **3.48** (**82**%)
		Sunny ext	L	2.11, 16.59 (2%)	0.3, 6.95 (50%)	0.35, 8.14 (39%)	**0.28**, **6.67** (**54**%)
SUNCG	Dense	Test sequence	1.07, 6.07 (17%)	0.75, 12.14 (13%)	N/A	N/A	**0.15**, **5.69** (**60**%)
	Sparse pos		1.21, 9.16 (7%)	1.14, 19.76 (2%)	N/A	N/A	**0.46**, **4.30** (**51**%)
	Sparse rot		1.24, 12.36 (2%)	0.89, 19.76 (6%)	N/A	N/A	**0.20**, **5.15** (**57**%)
	Sparse pos-rot		1.62, 20.16 (0%)	1.51, 24.51 (1%)	N/A	N/A	**0.72**, **18.58** (**19**%)

As can be seen, the challenging radiometric conditions in the COLD-Freiburg database caused the GPPF method to lose tracking, while the PRP estimator achieved low accuracy as a result of not exploiting the trajectory sequentiality, performing PR at every time-step instead. In turn, the solutions based on Network flow provide very accurate estimations in general, with the best results achieved by the uniformly sampled map under favorable conditions (i.e., *night* and *cloudy* sequences) and slightly worse in the case of severe appearance changes (i.e., *sunny* sequences). Our proposal, in contrast, demonstrates providing consistent results regardless of the appearance settings, achieving similar results to the Network flow solution in favorable conditions and outperforming all other methods under challenging radiometric settings.

In the case of the SUNCG dataset, the formulation proposed by the Network flow is incompatible with grid maps covering multiple rotations at the same location, as they are conceived to work only with positions. In turn, PRP and GPPF obtain low performance, even worsened in subsampled maps, while our proposal achieves the best results.

Despite its similarity with our approach, GPPF has shown to be unable to achieve robust localization due to the abovementioned issues: (i) deficiencies from considering a single pose–descriptor mapping for the whole environment, (ii) the absence of a relocalization process, and (iii) its limitation to environments without radiometric changes.

In conclusion, the presented comparison proves that these state-of-the-art localization methods based on appearance cannot provide both consistent and accurate localization estimations while operating within maps of diverse nature and captured under different appearance conditions. Our method, in turn, achieves higher performance in these conditions in terms of precision and robustness, showing notable results given its pure appearance nature.

5. Conclusions

We have presented a system for appearance-based robot localization that provides accurate, continuous pose estimations for camera navigation within a 2D environment under diverse radiometric conditions. Our proposal relies on the assumption that image global descriptors form a manifold articulated by the camera pose that adequately approximates the Image Manifold. This way, we gather pose–descriptor pairs from a lightweight map in order to create locally smooth regions called *Patches of Smooth Appearance Change* (PSACs) that shape, piecewise, the Descriptor Manifold geometry. Additionally, we robustly deal with appearance changes by modeling the descriptor variations under a white noise distribution.

We implemented a sequential camera tracking system built upon a Gaussian Process Particle Filter, which allows for multihypothesis pose estimation. Thus, our system optimizes a specific GP for each PSAC, subsequently being employed to define a local observation model of the descriptor for the Particle Filter. Furthermore, our method includes a relocalization process based on PR in case of tracking loss.

A first set of experiments has shown our proposal's error baseline in different environments and for a selection of holistic descriptors, revealing the most suitable configuration for our system. Finally, we have presented a comprehensive evaluation of the localization performance, showing that our approach outperforms state-of-the-art appearance-based localization methods in both tracking accuracy and robustness, even using images with challenging illuminations, yielding median errors below 0.3 m and 6°. Consequently, we have proven that pure appearance-based systems can produce continuous estimations with promising results in terms of accuracy, while working with lightweight maps and achieving robustness under strong appearance changes.

Future work includes research about (i) building the appearance map in an optimal way and wisely selecting where to sample the Descriptor Manifold; (ii) the design of a novel holistic descriptor that is more adequate to perform pose regression while keeping high invariance to radiometric changes.

Author Contributions: Conceptualization, A.J., F.-A.M., and J.G.-J.; methodology, A.J.; software, A.J.; validation, A.J.; investigation, A.J.; writing—original draft preparation, A.J., F.-A.M., and J.G.-J.; writing—review and editing, A.J., F.-A.M., and J.G.-J.; supervision, F.-A.M. and J.G.-J.; funding acquisition, J.G.-J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by: Government of Spain grant number FPU17/04512; by the "I Plan Propio de Investigación, Transferencia y Divulgación Científica" of the University of Málaga; and under projects ARPEGGIO (PID2020-117057) and WISER (DPI2017-84827-R) financed by the Government of Spain and European Regional Development's funds (FEDER).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal used for this research.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

PSAC	Patch of Smooth Appearance Change
PR	Place Recognition
GP	Gaussian Process
GPPF	Gaussian Process Particle Filter
PF	Particle Filter
SLAM	Simultaneous Localization and Mapping
IM	Image Manifold
DM	Descriptor Manifold
DL	Deep Learning
KP	Key-pair
CS	Constant Sampling

References

1. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.
2. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
3. Mur-Artal, R.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [CrossRef]
4. Sattler, T.; Leibe, B.; Kobbelt, L. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1744–1756.
5. Gomez-Ojeda, R.; Moreno, F.A.; Zuñiga-Noël, D.; Scaramuzza, D.; Gonzalez-Jimenez, J. PL-SLAM: A stereo SLAM system through the combination of points and line segments. *IEEE Trans. Robot.* **2019**, *35*, 734–746. [CrossRef]
6. Sattler, T.; Torii, A.; Sivic, J.; Pollefeys, M.; Taira, H.; Okutomi, M.; Pajdla, T. Are large-scale 3d models really necessary for accurate visual localization? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1637–1646.
7. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
8. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5297–5307.
9. Cummins, M.; Newman, P. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *Int. J. Robot. Res.* **2008**, *27*, 647–665. [CrossRef]
10. Milford, M.J.; Wyeth, G.F. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 1643–1649.
11. Pepperell, E.; Corke, P.I.; Milford, M.J. All-environment visual place recognition with SMART. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 1612–1618.
12. Radenović, F.; Tolias, G.; Chum, O. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1655–1668. [CrossRef] [PubMed]
13. Crowley, J.L.; Pourraz, F. Continuity properties of the appearance manifold for mobile robot position estimation. *Image Vis. Comput.* **2001**, *19*, 741–752. [CrossRef]
14. Ham, J.; Lin, Y.; Lee, D.D. Learning nonlinear appearance manifolds for robot localization. In Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2–6 August 2005; pp. 2971–2976.
15. He, X.; Zemel, R.S.; Mnih, V. Topological map learning from outdoor image sequences. *J. Field Robot.* **2006**, *23*, 1091–1104. [CrossRef]
16. Gomez-Ojeda, R.; Lopez-Antequera, M.; Petkov, N.; Gonzalez-Jimenez, J. Training a convolutional neural network for appearance-invariant place recognition. *arXiv* **2015**, arXiv:1505.07428.
17. Ko, J.; Fox, D. GP-BayesFilters: Bayesian filtering using Gaussian process prediction and observation models. *Auton. Robot.* **2009**, *27*, 75–90. [CrossRef]
18. Lopez-Antequera, M.; Petkov, N.; Gonzalez-Jimenez, J. Image-based localization using Gaussian processes. In Proceedings of the 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Alcala de Henares, Spain, 4–7 October 2016; pp. 1–7.
19. Roweis, S.T.; Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290*, 2323–2326. [CrossRef]
20. Tenenbaum, J.B.; De Silva, V.; Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319–2323. [CrossRef] [PubMed]

21. Pless, R.; Souvenir, R. A survey of manifold learning for images. *IPSJ Trans. Comput. Vis. Appl.* **2009**, *1*, 83–94. [CrossRef]
22. Wakin, M.B.; Donoho, D.L.; Choi, H.; Baraniuk, R.G. The multiscale structure of non-differentiable image manifolds. In Proceedings of the Wavelets XI, Proceedings of the SPIE, San Diego, CA, USA, 17 September 2005; Volume 5914, p. 59141B.
23. Lopez-Antequera, M.; Gomez-Ojeda, R.; Petkov, N.; Gonzalez-Jimenez, J. Appearance-invariant place recognition by discriminatively training a convolutional neural network. *Pattern Recognit. Lett.* **2017**, *92*, 89–95. [CrossRef]
24. Thoma, J.; Paudel, D.P.; Chhatkuli, A.; Van Gool, L. Learning Condition Invariant Features for Retrieval-Based Localization from 1M Images. *arXiv* **2020**, arXiv:2008.12165.
25. Jaenal, A.; Zuñiga-Nöel, D.; Gomez-Ojeda, R.; Gonzalez-Jimenez, J. Improving Visual SLAM in Car-Navigated Urban Environments with Appearance Maps. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October–24 January 2021; pp. 4679–4685.
26. Lowry, S.; Sünderhauf, N.; Newman, P.; Leonard, J.J.; Cox, D.; Corke, P.; Milford, M.J. Visual place recognition: A survey. *IEEE Trans. Robot.* **2015**, *32*, 1–19. [CrossRef]
27. Lenc, K.; Vedaldi, A. Understanding image representations by measuring their equivariance and equivalence. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 991–999.
28. Jaenal, A.; Moreno, F.A.; Gonzalez-Jimenez, J. Experimental study of the suitability of CNN-based holistic descriptors for accurate visual localization. In Proceedings of the 2nd International Conference on Applications of Intelligent Systems, Las Palmas de Gran Canaria, Spain, 7–9 January 2019; pp. 1–6.
29. Sattler, T.; Weyand, T.; Leibe, B.; Kobbelt, L. Image Retrieval for Image-Based Localization Revisited. In Proceedings of the British Machine Vision Conference, Surrey, UK, 3–7 September 2012; Volume 1, p. 4.
30. Torii, A.; Arandjelovic, R.; Sivic, J.; Okutomi, M.; Pajdla, T. 24/7 place recognition by view synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1808–1817.
31. Laskar, Z.; Melekhov, I.; Kalia, S.; Kannala, J. Camera relocalization by computing pairwise relative poses using convolutional neural network. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 929–938.
32. Balntas, V.; Li, S.; Prisacariu, V. RelocNet: Continuous Metric Learning Relocalisation using Neural Nets. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 751–767.
33. Ding, M.; Wang, Z.; Sun, J.; Shi, J.; Luo, P. CamNet: Coarse-to-fine retrieval for camera re-localization. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 2871–2880.
34. Sünderhauf, N.; Neubert, P.; Protzel, P. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In Proceedings of the Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA), Karlsruhe, Germany, 6–10 May 2013.
35. Naseer, T.; Spinello, L.; Burgard, W.; Stachniss, C. Robust visual robot localization across seasons using network flows. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Quebec City, QC, Canada, 27–31 July 2014; pp. 2564–2570.
36. Naseer, T.; Burgard, W.; Stachniss, C. Robust visual localization across seasons. *IEEE Trans. Robot.* **2018**, *34*, 289–302. [CrossRef]
37. Thoma, J.; Paudel, D.P.; Chhatkuli, A.; Probst, T.; Gool, L.V. Mapping, localization and path planning for image-based navigation using visual features and map. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7383–7391.
38. Maddern, W.; Milford, M.; Wyeth, G. CAT-SLAM: Probabilistic localisation and mapping using a continuous appearance-based trajectory. *Int. J. Robot. Res.* **2012**, *31*, 429–451. [CrossRef]
39. Rasmussen, C.E. Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 63–71.
40. Huhle, B.; Schairer, T.; Schilling, A.; Straßer, W. Learning to localize with Gaussian process regression on omnidirectional image data. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 5208–5213.
41. Schairer, T.; Huhle, B.; Vorst, P.; Schilling, A.; Straßer, W. Visual mapping with uncertainty for correspondence-free localization using Gaussian process regression. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 25–30 September 2011; pp. 4229–4235.
42. Lopez-Antequera, M.; Petkov, N.; Gonzalez-Jimenez, J. City-scale continuous visual localization. In Proceedings of the 2017 European Conference on Mobile Robots (ECMR), Paris, France, 6–8 September 2017; pp. 1–6.
43. Doucet, A.; de Freitas, N.; Gordon, N. *Sequential Monte Carlo Methods in Practice*; Statistics for Engineering and Information Science; Springer Science & Business Media: New York, NY, USA, 2001.
44. Blanco, J.L. *A Tutorial on SE(3) Transformation Parameterizations and On-Manifold Optimization*; Technical Report; Universidad de Málaga: Málaga, Spain, 2010; Volume 3, p. 6.
45. Pronobis, A.; Caputo, B. COLD: The CoSy localization database. *Int. J. Robot. Res.* **2009**, *28*, 588–594. [CrossRef]
46. Song, S.; Yu, F.; Zeng, A.; Chang, A.X.; Savva, M.; Funkhouser, T. Semantic scene completion from a single depth image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1746–1754.

47. Wu, Y.; Wu, Y.; Gkioxari, G.; Tian, Y. Building generalizable agents with a realistic and rich 3D environment. *arXiv* **2018**, arXiv:1801.02209.
48. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
50. GPy. GPy: A Gaussian Process Framework in Python. Since 2012. Available online: http://github.com/SheffieldML/GPy (accessed on 30 March 2021).