Unsupervised appearance map abstraction for indoor Visual Place Recognition with mobile robots

Alberto Jaenal Francisco-Angel Moreno Javier Gonzalez-Jimenez

***Abstract*— Visual Place Recognition (VPR), the task of identifying the place where an image has been taken from, is at the core of important robotic problems as relocalization, loop-closure detection or topological navigation. Even for indoors, the focus of this work, VPR is challenging for a number of reasons, including real-time performance when dealing with large image databases ($\sim 10^4$) (probably captured by different robots), or the avoidance of Perceptual Aliasing in environments with repetitive structures and scenes.**

In this paper, we tackle these issues by proposing an off-line mapping technique that abstracts a dense database of georeferenced images without particular order into a Multivariate Gaussian Mixture Model, by creating soft clusters in terms of their similarity in both pose and appearance. This abstract representation is obtained through an Expectation-Maximization algorithm and plays the role of a simplified map. Since querying this map yields a probability of being in a cluster, we exploit this "belief" within a Bayesian filter that regards previous query images and a topological map between clusters to perform more robust VPR.

We evaluate our proposal in two different indoor datasets, demonstrating comparable VPR precision to querying the full database while incurring in shorter query times and handling Perceptual Aliasing for sequential navigation.

***Index Terms*— Place Recognition, Map Abstraction, Appearance-based localization**

I. Introduction

Visual Place Recognition (VPR) [1], [2] aims to detect the most similar place to a certain query image, given a map consisting of a generally large database of georeferenced images. This task has received increasing attention during the last decades in the robotic community, due to its involvement in important areas as loop-closure detection, re-localization, or topological navigation. For such tasks, the VPR database is built from geo-tagged images collected during several robot navigations that are encoded with some global descriptor [3], [4], either as a sequence [5], or as a set of unordered elements [6]. This database is treated as an *Appearance Map* (AM) of the environment. In the case of indoors, where the robot may revisit multiple times some parts of the environment, the AM will typically include repeated views. Not only does this not contribute any meaningful information to the map, but it increases its size to typically tens of thousands of images.

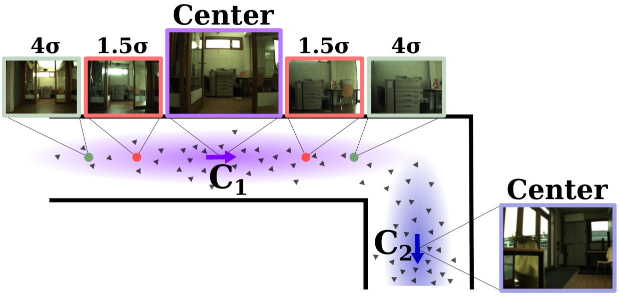

Fig. 1: Our work aims to abstract unordered georreferenced images (black triangles) into *clusters* C_j defined as multivariate Gaussian distributions (colored ellipses). These distributions represent spatial regions with visual appearance resemblance that can be interpreted as *places*.

Commonly, VPR is addressed on such AMs as an Image Retrieval (IR) problem that searches the Nearest Neighbors (NNs) of the query image according to a certain appearance similarity measure (e.g. Euclidean distance between descriptors). Then, the procedure yields an estimated location from the k most similar elements in the database.

This IR approach presents the following limitations:

- It usually follows a similarity criteria for categorization using only the appearance, disregarding the spatial aspect of VPR, hence not being able to deal with Perceptual Aliasing (i.e. places distant in pose but sharing similar appearance). This subsequently leads to incorrect pose estimations. In traditional VPR, this issue is typically solved by using additional topology from sequential databases [7], [8], [9], unavailable for unordered maps.
- The selection of the NNs follows a *hard* classification approach, as no information about their confidence is provided. This makes IR more difficult to recover from incorrect query results, as well as unable to be included in probabilistic frameworks.
- Querying large databases becomes highly time-consuming, often hindering real-time operation as required by mobile robotics applications.
- The result is a set of discrete, unrelated candidates where reliable pose interpolation is not possible, so post-processing [10], [11], [12] is commonly required to obtain a refined estimation for the image pose.

Focusing on performing robust VPR in indoors with mobile robots, we propose in this work to off-line abstract the information stored at large databases of geo-tagged images

This work has been funded by the Government of Spain in part under grant FPU17/04512, in part under the research project *ARPEGGIO* (PID2020-117057GB-I00), funded by the European H2020 program. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal used for this research.

The authors are members of the Machine Perception and Intelligent Robotics Group (MAPIR-UMA), within the Malaga Institute for Mechatronics Engineering & Cyber-Physical Systems (IMECH.UMA), University of Malaga, Spain. {ajaenal, famoreno, javiergonzalez}@uma.es

without any established order into a set of clusters with associated probabilistic information (as depicted in Fig. 1). Our proposal represents the map as a Multivariate Gaussian Mixture Model (MGMM), grouping images that are both similar in appearance and pose over a new, joint pose-appearance space. The parameters of each distribution are unsupervisedly estimated through an *Expectation-Maximization* (EM) formulation. Thus, each cluster represents a *place* that can be identified by a VPR system.

This approach presents a series of advantages to cope with the aforementioned IR limitations:

- Creating clusters by taking into account not only the image appearance but also their poses allows us to robustly handle Perceptual Aliasing during the clusterization.
- Our approach eliminates redundant information from the map, since each group of images is represented by the probability distribution that simultaneously best fits their pose and appearance (see Fig. 1). This way, given a query image, our proposal yields a probability value for each cluster, avoiding hard classification and allowing for multi-hypothesis instead.
- Since the number of clusters is significantly lower than that of elements in the original database, we can maintain adequate VPR precision while incurring in much shorter query times.
- Despite working with unordered image databases[1], we propose a topological model based on the pose information of the images to generate transitions between the clusters. This can be further exploited as topology to improve VPR and to avoid PA effects during localization. On the contrary, typical VPR approaches would require sequential databases to obtain such improvement.

To demonstrate this, we validate our proposal in two different indoor datasets employing three different state-of-the-art image global descriptors, in order to, first, create an abstracted map with an associated topology, and then perform probabilistic VPR within it for a sequence of query images. The results show that our approach naturally deals with PA and manages multiple-hypothesis in the pose estimation, effectively converging to the actual cluster in short time. The performance of our VPR proposal, in terms of precision-recall, is comparable to querying the full database with a threshold of $(5.0m, 10^{\circ})$ while incurring in a fraction of the computational cost, and achieves similar performance than other state-of-the-art VPR methods. Our code for map abstraction is available[2], as well as a demonstration video[3] .

II. Related works

Map simplification aims to retain the most representative subset of samples from a large map, removing elements that do not contribute to the localization because of their redundancy or lack of distinctiveness. The topic has been thoroughly studied in Visual Localization and SLAM, where matching 2D local features against large 3D models becomes expensive and inefficient [13], proposing some solutions as 3D model compression [14], [15], prioritized searches [16], matching constraints [17] or feature temporal modeling [18].

On the other hand, while geo-tagged image databases for VPR typically represent large environments, the main efforts during the last years for better VPR scalability have focused on optimizing the search either (i) by reducing the size of the descriptor through appearance-aware representations as quantization [19], [20] and binarization [21], [22]; or (ii) by replacing NNs approaches with more efficient querying techniques such as hashing [8], [9]. In this work, we propose to improve VPR scalability on large databases through map abstraction, that is, modelling the pose and appearance of the map elements so that redundancies can be removed. Furthermore, the probabilistic nature of such abstraction allows us to provide additional information to the map.

Commonly, sparse appearance representations are obtained by applying uniform sampling to single-sequence databases according to thresholds in pose [7], [23], [24], which produces an ordered set of key-samples. Aiming to improve such selection, [23] takes into account both appearance and position in a network flow formulation, while SeqNet [25] introduces descriptors for short subsequences that are employed in a hierarchical VPR framework. Some authors have studied the abstraction of multi-sequence databases: in [26] coresets are employed for an hierarchical summarization of the environment; Vysotka et al. [27] employ an association graph to deal with retrieval and relocalization and [28] propose smooth interpolation areas for accurate localization.

In contrast, our approach is able to handle unordered databases without requiring prior topology and produces a map abstraction in the form of distributions in pose and appearance representing *places* in the original map.

III. Appearance map abstraction and Visual Place Recognition

In this section, we first provide a description of the proposed off-line map abstraction approach, grounded on the MGMM parameter estimation through the *Expectation-Maximization* (EM) algorithm. Then, we describe a probabilistic Visual Place Recognition pipeline that builds upon the set of clusters estimated by the EM algorithm to improve its robustness against Perceptual Aliasing

A. Appearance map and combined space

Let us first formally define an *Appearance Map* (AM)

$$\mathcal{M} = \{(\mathbf{p}^i, \mathbf{d}^i)\}_{i=1}^{N}, \tag{1}$$

as an unordered set of N pairs, each one formed by a D-dimensional global image descriptor $\mathbf{d}^i \in \mathbb{R}^D$ and the 2D pose $\mathbf{p}^i \in \mathrm{SE}(2)$ from where the image was taken.

Since our key motivation is to simultaneously exploit the pose and appearance coherence, we build a joint pose-descriptor space $\mathcal{E} = \mathrm{SE}(2) \oplus \mathbb{R}^D$ whose elements $\mathcal{X} = \{\mathbf{x}^i\}_{i=1}^{N}$ result from the concatenation of the map samples

[1] Our proposal does not rely on any pre-established order of the input data, i.e. sequences.

[2] https://github.com/AlbertoJaenal/MapAbstractionVPR

[3] https://youtu.be/i6oXaAKcEII

components: $\mathbf{x}^i = \left[\mathbf{p}^i \ \mathbf{d}^i\right]^\intercal \in \mathcal{E}$. Our proposal performs EM-based clustering in this combined space.

B. Expectation Maximization

The Expectation-Maximization (EM) algorithm [29] aims to estimate the parameters of a distribution for some known input data, applying Maximum Likelihood Estimation over a given likelihood function.

Specifically, for a given input AM $\mathcal{M}$, our EM formulation finds the parameters of a set of M clusters $\mathcal{C} = \{C_j\}_{j=1}^{M}$. Each cluster $C_j = \left(\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), P(C_j)\right)$ represents a *place* in $\mathcal{M}$ and is composed of: the mean $\boldsymbol{\mu}_j \in \mathcal{E}$ and covariance matrix $\boldsymbol{\Sigma}_j \in \mathcal{E} \times \mathcal{E}$ of a multivariate Gaussian distribution lying in the joint space $\mathcal{E}$, and its prior probability $P(C_j)$.

The EM algorithm iteratively applies two steps: (i) **Expectation**, where the posterior probabilities of the clusters are estimated for the given data samples:

$$P\Big(C_j|\mathbf{x}^i\Big) = \frac{P\Big(C_j\Big)P\Big(\mathbf{x}^i|C_j\Big)}{\sum_{j=1}^{N} P\Big(C_j\Big)P\Big(\mathbf{x}^i|C_j\Big)}, \tag{2}$$

and (ii) **Maximization**, where the cluster parameters (i.e. $P\Big(C_j\Big)$, μ_j, and Σ_j) are updated to maximize the probability of the data. These steps are repeated until convergence to obtain an *abstracted* map, that is, the optimal set of clusters that best represents the data:

$$\mathcal{C}^* = \arg\max_{\mathcal{C}} P\big(\mathcal{X}|\mathcal{C}\big). \tag{3}$$

In (2), the likelihood of a sample pair $\mathbf{x}^i$ belonging to the cluster C_j is defined as a Gaussian density function on $\mathcal{E}$:

$$P\big(\mathbf{x}^i|C_j\big) \sim \mathcal{N}\big(\mathbf{x}^i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\big). \tag{4}$$

However, the high dimensionality of state-of-the-art image descriptors (typically $D > 1000$) causes $\mathcal{E}$ to become an highly undersampled space, i.e. the number of input samples $|\mathcal{X}|$ for the EM is significantly scarce at such dimensions.

To alleviate this situation, we propose to approximate (4) adopting the next assumptions about the covariance matrix:

- The pose and descriptor of an image are conditionally independent given a certain cluster ($\mathbf{p}^i \perp\!\!\!\perp \mathbf{d}^i \mid C_j$).
- All the components of the descriptor are independent between them and share the same variance value σ_d^2.

Thus, we can rewrite (4) as:

$$\begin{aligned} P\Big(\mathbf{x}^i|C_j\Big) &\sim \mathcal{N}\left((\mathbf{p}^i, \mathbf{d}^i)\middle| \begin{bmatrix}\boldsymbol{\mu}_j^p \\ \boldsymbol{\mu}_j^d\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_j^p & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_j^d\end{bmatrix}\right) = \\ &= \mathcal{N}\left(\mathbf{p}^i|\boldsymbol{\mu}_j^p, \boldsymbol{\Sigma}_j^p\right)\mathcal{N}\left(\mathbf{d}^i|\boldsymbol{\mu}_j^d, \boldsymbol{\Sigma}_j^d\right), \end{aligned} \tag{5}$$

being $(\boldsymbol{\mu}_j^d, \boldsymbol{\Sigma}_j^d = \sigma_j^{d^2} \cdot \mathbf{I}^D)$ the mean and simplified covariance matrix of the descriptor distribution and $(\boldsymbol{\mu}_j^p \in SE(2), \boldsymbol{\Sigma}_j^p \in \mathfrak{se}(2) \times \mathfrak{se}(2))$ the distribution parameters for the pose. Note that, to consistently handle rotations, the pose distribution is defined on the tangent space $\mathfrak{se}(2)$. Thus, the pose probability of a sample at such distribution is given by:

$$\begin{aligned} &\mathcal{N}\left(\mathbf{p}^i|\boldsymbol{\mu}_j^p, \boldsymbol{\Sigma}_j^p\right) = \\ &= \frac{1}{\sqrt{(2\pi)^3\left|\boldsymbol{\Sigma}_j^p\right|}} \exp\left(-\frac{1}{2}\xi_j^{i\intercal}\left(\boldsymbol{\Sigma}_j^p\right)^{-1}\xi_j^i\right), \end{aligned} \tag{6}$$

where the pose twist [30] from the distribution mean to the sample is denoted by $\xi_j^i = \log(\boldsymbol{\mu}_j^p \ominus \mathbf{p}^i) \in \mathfrak{se}(2)$.

However, the appearance term in (5) still suffers from the impact of its high dimensionality in the normalization factor of the Gaussian density function. To mitigate this, we approximate it as the following univariate distribution:

$$\begin{aligned} &\mathcal{N}(\mathbf{d}^i|\boldsymbol{\mu}_j^d, \boldsymbol{\Sigma}_j^d) = P(\mathbf{d}^i|C_j) = \\ &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}_j{}^d|}} \exp\left(-\frac{1}{2}(\mathbf{d}^i - \boldsymbol{\mu}_j^d)^\intercal\left(\boldsymbol{\Sigma}_j^d\right)^{-1}(\mathbf{d}^i - \boldsymbol{\mu}_j^d)\right) \\ &\approx \frac{1}{\sqrt{2\pi\sigma_j^{d^2}}} \exp\left(-\frac{1}{2}\frac{||\mathbf{d}^i - \boldsymbol{\mu}_j^d||_2^2}{\sigma_j^{d^2}}\right). \end{aligned} \tag{7}$$

The appearance variance for each cluster $\sigma_j^{d^2}$ is estimated in the maximization step as:

$$\sigma_j^{d^2} = \frac{1}{\sum_i P\left(C_j|\mathbf{x}^i\right)} \sum_i P\left(C_j|\mathbf{x}^i\right)||\mathbf{d}^i - \boldsymbol{\mu}_j^d||_2^2. \tag{8}$$

C. Topological model

The topological model of the abstracted map $\mathcal{C}^*$ refers to the probability of the robot moving between two clusters $P(C_j|C_k)$. Such probability is computed according to the pose term of each cluster:

$$P(C_j|C_k) \propto \begin{cases} \mathcal{N}\left(\boldsymbol{\mu}_j^p|\boldsymbol{\mu}_k^p, \boldsymbol{\Sigma}_k^p\right) + \mathcal{N}\left(\boldsymbol{\mu}_k^p|\boldsymbol{\mu}_j^p, \boldsymbol{\Sigma}_j^p\right) & \text{if } j \neq k \\ \max_k P(C_j|C_k) & \text{otherwise} \end{cases} \tag{9}$$

The transition probabilities are subsequently normalized for each cluster.

D. Probabilistic Visual Place Recognition

Once the abstracted map $\mathcal{C}^*$ is built, performing probabilistic VPR (pVPR) within it translates into estimating the probability $P(C_j|\mathbf{d}^q)$ of each cluster to contain a particular query descriptor $\mathbf{d}^q$. The pVPR output probabilities correspond to *places*, i.e. pose distributions $\mathcal{N}\big(\boldsymbol{\mu}_j^p, \boldsymbol{\Sigma}_j^p\big)$ of the most probable clusters. Note that this implies using only appearance information as input, leaving the pose component of the joint space $\mathcal{E}$ unobservable. Consequently, only the cluster information related to the appearance can be taken into account for querying. In fact, this may incur in Perceptual Aliasing when querying the map. This issue is handled by defining a topology over the abstracted map, which, along with the probabilistic nature of $\mathcal{C}*$, allow us to disambiguate the place estimation through recurrence.

Commonly, robots navigate indoors equipped with a camera while gathering a sequence of images taken at T timesteps, which in turn are transformed into appearance descriptors: $\mathbf{d}^{q,T} = [\mathbf{d}_1^q, \mathbf{d}_2^q, ..., \mathbf{d}_T^q]$. We assume that the robot never leaves the space covered by any of the clusters in $\mathcal{C}^*$.

In this situation, the probability of the robot being in C_j at a time t is expressed with the topological filter:

$$P(C_j|\mathbf{d}^{q,t}) = P(\mathbf{d}_t^q|C_j) \sum_k P(C_j|C_k) P(C_k|\mathbf{d}^{q,t-1}) \quad (10)$$

IV. Experimental evaluation

This section describes the experiments conducted to assess the outcome of our proposal for map abstraction and VPR.

A. Datasets

We employed two publicly available datasets captured at indoor environments by a robot as unordered sets of descriptor-pose pairs, with available revisits under different appearance conditions:

- The COLD database [5] consists of images from a real-world office (Freiburg, *part A*) gathered at 5 Hz. The input database for the map abstraction includes all available images from cloudy days (*Cloudy-database*), totalling $\sim$ 13k unordered images from six different sequences (each one visiting different parts of the environment). The evaluation comprehends *Seq2_night1*, a full navigation of the environment under artificial illumination (*Artificial lights*), and *Seq1_sunny1*, a partial visitation on a sunny day (*Sunny-partial*), with more challenging appearance changes.
- The Robot at Virtual Home (R@VH) dataset [31] provides images within a simulated house (*Home11*) at 30 Hz. In this case, the database consists of $\sim$ 30k images from a random navigation with artificial illumination at night that contains multiple revisits. The evaluation sequences includes a single navigation through the whole house under three different settings for the artificial illumination: completely on (*Artificial lights*), with similar appearance; randomly activated (*Random lights*), which is more challenging; and completely off at dusk (*Dark*), really challenging.

Based on [24], we extracted from each evaluation sequence 100 subsequences composed of 200 images gathered every $0.15m$.

Regarding the appearance representation, we employ three global image descriptors: (i) 4096-sized NetVLAD descriptor [4][4]; (ii) 2048-sized Resnet-101 Generalized Mean (GeM) descriptor from [33]; and (iii) a Bag of Words (BoW) descriptor [3] built from ORB features, with vocabulary trained in [34] accumulated into 1024 bins.

B. Map abstraction setup

One of the key aspects for the EM algorithm to converge is the selection of the initial set of clusters.

This initialization is accomplished by applying the *K-Means* (KM) clusterization method to the SE(2) poses of the input samples, represented as points in the form $(x, y, \theta) \in \mathbb{R}^3$. The decision on the number of clusters M, however, required a careful study that consisted on the evaluation of several KM outputs with varying M according to the Davies-Bouldin (DB) index [35]. This index rewards clusterizations with dense and separate clusters, being optimal when minimizing:

$$DB = \frac{1}{M} \sum_{i=1}^{M} \max_{j \neq i} \left\{ \frac{\bar{d}_i + \bar{d}_j}{d_{i,j}} \right\}, \quad (11)$$

with the average pose radius of the i-th cluster to its centroid $\boldsymbol{\mu}_i^p$ denoted by $\bar{d}_j = \sqrt{\frac{1}{T_i} \sum_{k=1}^{T_i} d_{SE(2)}(\mathbf{x}^k, \boldsymbol{\mu}_i^p)^2}$; and $d_{i,j} = d_{SE(2)}(\boldsymbol{\mu}_i^p, \boldsymbol{\mu}_j^p)$ the distance between the centroids of the i-th and j-th clusters. The metric $d_{SE(2)}(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{t}_1 - \mathbf{t}_2||_2 + |\log(\mathbf{R}_1^\intercal \mathbf{R}_2)|)$ ensures a consistent distance on SE(2). The results of this study can be seen in Fig. 2, which led us to select the M values with lowest DB index: 35, 65 and 90 for COLD and 38, 74 and 98 for R@VH.

Afterwards, the EM cluster means $(\boldsymbol{\mu}_j)$ were initialized using the nearest database samples in terms of pose to the centroids of the resulting KM clusters. The initial EM covariances for poses and descriptors ($\boldsymbol{\Sigma}_j^p$ and $\boldsymbol{\Sigma}_j^d$) were obtained from the samples forming each resulting KM cluster.

Finally, and regarding the map topology, we found that the clusters tended to group samples with different positions while keeping similar rotations, hence yielding small covariance values for the orientation (Fig. 2). This happens especially in long corridors, where the robot movement is mostly linear. When using this covariance into the transition model in (6), the orientation term becomes excessively large, only connecting clusters with similar rotations. In order to avoid such situations, we have added to the rotational part of the covariance matrix Σ_k^p in (6) an additional value ${\sigma^*}_\theta^2 = 1$.

C. Map abstraction validation

This section evaluates the representativeness of the created map with respect to the environment, measured as the mean appearance distance between a sample and the samples nearby within a given threshold.

For an abstracted map, we consider that a sample q falls within a cluster c if its pose lies in the cluster pose distribution with a 99.9% confidence. In the SE(2) 3-dimensional space, this equals to remain in a region within $\boldsymbol{\mu^p} \pm 4\boldsymbol{\Sigma^p}$ [36]. Thus, the Mahalanobis distance $\mathcal{D}_{c,q}$ between the query and the cluster must be less than 4.

$$\mathcal{D}_{c,q} = \sqrt{{\xi_{c,q}^p}^\intercal \left(\boldsymbol{\Sigma}_c^p\right)^{-1} \xi_{c,q}^p}, \quad (12)$$

with $\xi_{c,q}^p = \log(\boldsymbol{\mu}_c^p \ominus \mathbf{p}^q)$.

Table I compares, for the COLD database, the representativeness for three common pose thresholds in IR and the described confidence interval for three abstracted maps with different M values. In the case of the IR databases, we computed (12) for all the database samples within the specified threshold, while in the case of the abstracted map is only computed for each cluster. The results show that our proposal groups samples with an appearance similarity as good as that for the two most restrictive IR pose thresholds.

[4] Whose implementation is available at [32].

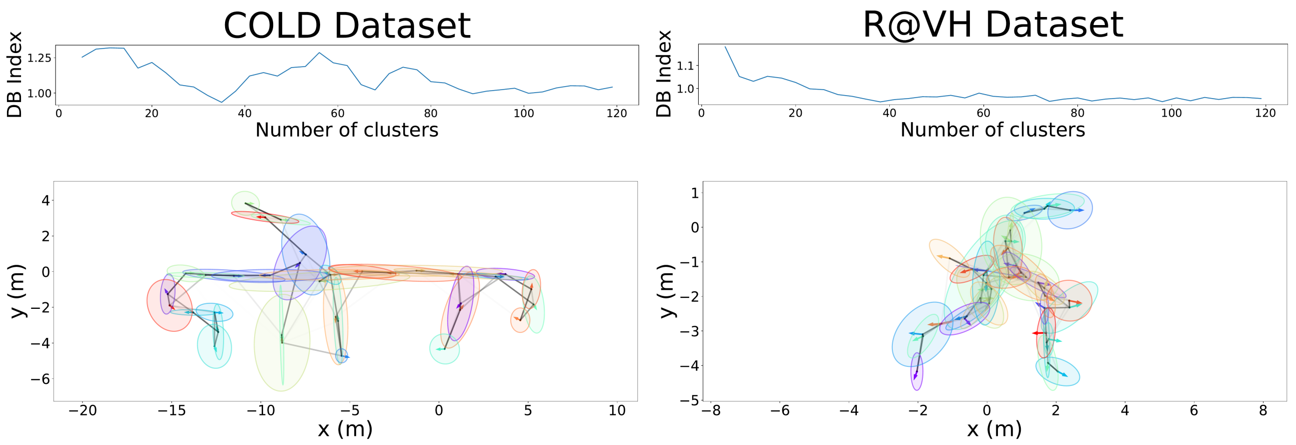

Fig. 2: First row: preliminar study of the Davies-Bouldin (DB) Index (11) for K-Means maps with varying number of clusters. Second row: example of two abstracted maps with minimum DB in $\mathbb{R}^2$ ($M = 35$ for COLD and $M = 38$ for R@VH), depicting the means (arrows), covariances (color blobs) and topology (black lines) for both datasets

TABLE I: MEAN DESCRIPTOR DISTANCE FOR COLD

Database	*NetVLAD*	*ImRet*	*ORB+BoW*
Full database $0.25m, 2°$	0.6408	0.3649	0.8245
Full database $0.50m, 5°$	0.8454	0.5140	0.9891
Full database $5.0m, 10°$	1.1497	0.8727	1.1457
EM $M = 35$	0.7903	0.5751	0.7921
EM $M = 65$	0.7510	0.5312	0.7803
EM $M = 90$	0.7459	0.5067	0.7725

D. Visual Place Recognition

In this section, we discuss the VPR outcome for our proposal, compared to conducting IR on the full database. This evaluation is provided in terms of (i) precision, (ii) query time, (iii) robustness against Perceptual Aliasing (PA) and (iv) precision in single sequence databases. However, we need to clarify that, due to the probabilistic nature of our abstracted map, it cannot be directly compared with standard IR methods when assessing VPR performance.

Typical VPR evaluation on geo-referenced databases [2], [37] relies on the pose error between the retrieved image and the query [8], [23], [24]. Thus, a query is considered to be correctly localized if such error falls below a certain tolerance. In contrast, our pVPR proposal defines an abstracted map as a collection of pose-appeaeance probability distributions that span across the sampled joint space of the original database, represented by their means and covariances. Consequently, the pose error between the query image and the mean of the selected distribution does not properly represent the precision of our system, as the clusters will represent spatial regions of varying sizes. In that manner, we consider a query to be correctly localized if its pose falls under the 99.9% confidence interval for the retrieved cluster.

1) Precision comparative: First, we have validated the VPR precision performance by means of the precision-recall metric for an evaluation sequence. Table II depicts the Area Under the Curve (AUC) for the precision-recall metric of both approaches for each dataset and descriptor, including three different numbers of clusters for each abstracted map. The results seem to show consistency between the DB index for each M and the AUC value, especially for the COLD database. For the R@VH dataset, though, the drop in precision for high M values might be caused for the excessive number of clusters in a relatively small environment, hence leading to excessive PA. In any case, the results show comparable precision for both datasets between our approach with $M \approx 35$ and a standard IR procedure on the full dataset with the threshold of $(5m, 10°)$.

TABLE II: PLACE RECOGNITION PRECISION (AUC).

	Database	*NetVLAD*	*ImRet*	*ORB+BoW*
COLD	*Full database* $0.25m, 2°$	0.2940	0.2635	0.0938
	Full database $0.50m, 5°$	0.6396	0.6585	0.3158
	Full database $5.0m, 10°$	0.9297	0.9329	0.5646
	pVPR (EM) $M = 35$	0.9340	0.8302	0.5702
	pVPR (EM) $M = 65$	0.6432	0.6175	0.5325
	pVPR (EM) $M = 90$	0.8682	0.7404	0.5227
R@VH	*Full database* $0.25m, 2°$	0.2406	0.2764	0.0665
	Full database $0.50m, 5°$	0.6730	0.7709	0.2579
	Full database $5.0m, 10°$	0.7916	0.9435	0.3121
	pVPR (EM) $M = 38$	1.0000	1.0000	0.8152
	pVPR (EM) $M = 74$	0.9638	0.5473	0.4235
	pVPR (EM) $M = 98$	0.5762	0.3400	0.3470

2) Time performance: We have measured the computational time spent for a single query by each of the previously assessed methods. The experimental evaluation is carried out with an Intel Core i7-6700K desktop computer with 16-GB RAM, employing the NumPy library. Note that the computational time does not include the descriptor extraction.

Table III compares the average time per step for the pVPR and for one IR query. Due to the sparsity of our abstracted map, pVPR query times becomes more than two orders of magnitude smaller for all the evaluated maps, proving its capacity for real time operation within unordered databases.

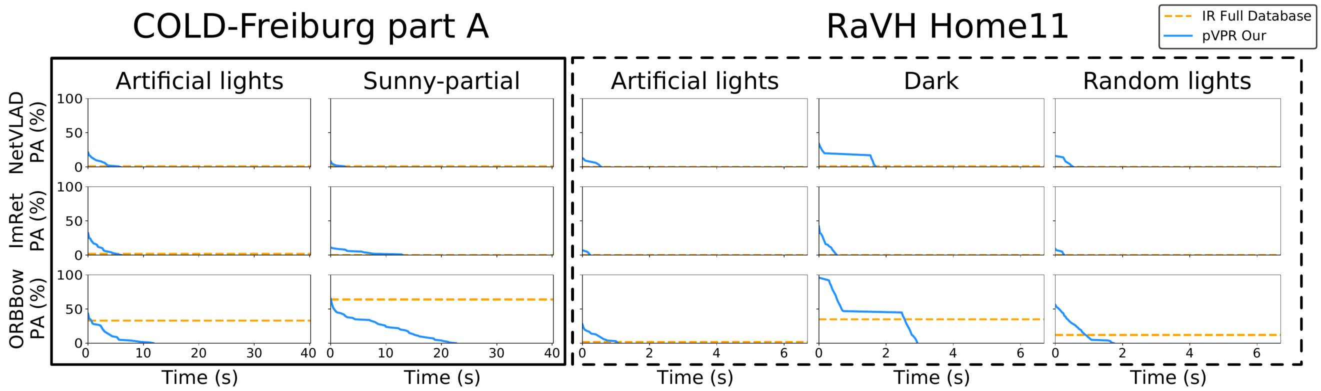

Fig. 3: Perceptual Aliasing (% of incorrectly localized queries) incurred during the localization for the different evaluation sequences on maps with $M \approx 35$, fixed for single-image IR. Note that the sampling frequency is different for each dataset.

TABLE III: MEAN QUERY TIME PER STEP (ms)

Database		*NetVLAD*	*ImRet*	*ORB+BoW*
COLD	*Full database*	160.0659	80.6049	40.6822
	pVPR (EM) $M = 35$	0.1476	0.0946	0.0708
	pVPR (EM) $M = 65$	0.2282	0.1297	0.0896
	pVPR (EM) $M = 90$	0.3773	0.1766	0.1100
R@VH	*Full database*	525.2183	255.3932	129.8930
	pVPR (EM) $M = 38$	0.1408	0.0908	0.0658
	pVPR (EM) $M = 74$	0.2562	0.1452	0.0988
	pVPR (EM) $M = 98$	0.9914	0.5202	0.4077

3) Perceptual Aliasing: As stated before, PA is one of the main challenges of VPR, where similar images but far in pose cause incorrect localization estimates. We aim to evaluate the robustness of our proposal against such circumstance compared to performing IR over the full original database.

For that purpose, we must first define the effect of Perceptual Aliasing over localization, i.e. when an estimate is incorrect. For that, we define the next thresholds: in the case of IR, we define a pose error of $(2.5m, 60°)$ while for our approach we keep the same threshold (out of the 99.9% confidence). Then, we measure, for all the sampled subsequences, the total percentage of incorrect estimations when applying IR and our pVPR filter on the maps of $M \approx 35$ with topology as described in Section III-D.

Fig. 3 depicts the time spent until convergence between both approaches for each dataset, descriptor and evaluation sequence. Note that, as IR is a single-shot method, the percentage of PA is constant over time. Besides, as the full database is dense in both datasets (see Section IV-A), the high sampling of the environment produces generally low PA percentage for the IR, although not zero. In this scenario, our proposal initial estimation is worse but, thanks to the constructed topology and the recursive estimation, it is able to quickly eliminate the effects of PA, eventually leading the pVPR localization to converge to the real cluster. It is important to remark here that the topology has been built from an initial set of unordered images, and no other sequential information has been employed during the map abstraction and topology construction.

Finally, in terms of time efficiency, and relying on the results shown in Section IV-D.2, our proposal requires smaller convergence time than a single IR query on the full database.

4) VPR on sequential databases: Typical VPR approaches localize on single-sequence databases, exploiting sequential topology to provide accurate estimates. In order to enable comparison, we abstracted a database composed only of the *Seq2_cloudy2* from the COLD Database. We compare the mean localization performance on both COLD evaluation sequences employing NetVLAD, using our pVPR localization on a map with $M = 35$ against IR and two online available state-of-the-art methods: the topological filter from [24] with NetVLAD descriptors and SeqSLAM[7] with images subsampled to 48×64. As a final note, take into account that the two later approaches exploit the sequentiality between the database elements, while our approach considers them as unconnected during the map abstraction.

TABLE IV: PLACE RECOGNITION PRECISION (AUC) (*COLD* DATASET & NETVLAD DESCRIPTOR)

Model	$0.25m, 2°$	$0.50m, 5°$	$5.0m, 10°$	P=99.9%
Single IR	0.0202	0.2828	0.7246	-
Top. Filt. [24]	0.1748	0.4032	0.8385	-
SeqSLAM [7]	0.0194	0.0269	0.0913	-
pVPR $M = 35$	-	-	-	0.5849

Table IV depicts the AUC score for each method, demonstrating that SeqSLAM is not able to perform in this database, while our approach obtains comparable or improved performance than the remaining methods with pose error thresholds of $(0.5m, 5°)$ and $(5.0m, 10°)$.

V. Conclusions and future work

In this work, we have presented a method that abstract dense, unordered image databases of indoor environments, by grouping similar images in both pose and appearance into soft clusters with associated probabilistic information in an off-line process. These clusters can be topologically connected and represent *places* in the map, defined as a Multivariate Gaussian Mixture Model in a joint pose-appearance space. This unsupervised clusterization process is based on the Expectation-Maximization algorithm.

Our method addresses Perceptual Aliasing during the map abstraction thanks to performing clusterization in the combined pose-appearance space. Besides, its effects in VPR are avoided by exploiting the probabilistic nature of the abstracted map and the created topology between clusters within a Bayesian localization filter for sequential queries.

We have determined the optimal number of clusters for each dataset, and have evaluated how the abstracted maps represent their environment in terms of appearance, compared with typical pose error thresholds. The probabilistic VPR filter has been evaluated in two indoor datasets, demonstrating comparable precision to IR over the full database in considerably shorter times. The filter also demonstrates to address Perceptual Aliasing for sequential data on the abstracted maps, besides achieving comparable performance to state-of-the-art methods on single sequence databases.

Future work includes: (i) to extend the applicability of this approach to outdoor environments, regarding more challenging conditions as SE(3) poses and lack of structure; (ii) studying different topological models, more suited to the abstracted representation; and (iii) a hierarchical localization model build upon the abstracted maps able to provide accurate geometrical pose estimations for the queries.

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[2] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, vol. 9, pp. 19 516–19 547, 2021.

[3] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, October 2012.

[4] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[5] A. Pronobis and B. Caputo, "Cold: The cosy localization database," *IJRR*, vol. 28, no. 5, pp. 588–594, 2009.

[6] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "InLoc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7199–7209.

[7] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *International Conference on Robotics and Automation*, 2012, pp. 1643–1649.

[8] O. Vysotska and C. Stachniss, "Relocalization under substantial appearance changes using hashing," in *Proceedings of the IROS Workshop on Planning, Perception and Navigation for Intelligent Vehicles, Vancouver, BC, Canada*, vol. 24, 2017.

[9] S. Garg and M. Milford, "Fast, compact and highly scalable visual place recognition through sequence-based matching of overloaded representations," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3341–3348.

[10] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *European Conference on Computer Vision*. Springer, 2020, pp. 726–743.

[11] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.

[12] A. Jaenal, D. Zuñiga-Nöel, R. Gomez-Ojeda, and J. Gonzalez-Jimenez, "Improving visual slam in car-navigated urban environments with appearance maps," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 4679–4685.

[13] A. Torii, H. Taira, J. Sivic, M. Pollefeys, M. Okutomi, T. Pajdla, and T. Sattler, "Are large-scale 3D models really necessary for accurate visual localization?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 814–829, 2019.

[14] S. Cao and N. Snavely, "Minimal scene descriptions from structure from motion models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 461–468.

[15] F. Camposeco, A. Cohen, M. Pollefeys, and T. Sattler, "Hybrid scene compression for visual localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[16] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, 2017.

[17] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[18] E. Johns and G.-Z. Yang, "Place recognition and online learning in dynamic scenes with spatio-temporal landmarks." in *BMVC*, 2011.

[19] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 1, pp. 117–128, 2010.

[20] Y. Kalantidis and Y. Avrithis, "Locally optimized product quantization for approximate nearest neighbor search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[21] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Towards life-long visual localization using an efficient matching of binary sequences from images," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 6328–6335.

[22] S. Lowry and H. Andreasson, "Lightweight, viewpoint-invariant visual place recognition in changing environments," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 957–964, 2018.

[23] J. Thoma, D. P. Paudel, A. Chhatkuli, T. Probst, and L. V. Gool, "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*, 2019, pp. 7383–7391.

[24] M. Xu, N. Sünderhauf, and M. Milford, "Probabilistic visual place recognition for hierarchical localization," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 311–318, 2020.

[25] S. Garg and M. Milford, "SeqNet: Learning descriptors for sequence-based hierarchical place recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4305–4312, 2021.

[26] M. Volkov, G. Rosman, D. Feldman, J. W. Fisher, and D. Rus, "Coresets for visual summarization with applications to loop closure," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 3638–3645.

[27] O. Vysotska and C. Stachniss, "Effective visual place recognition using multi-sequence maps," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1730–1736, 2019.

[28] A. Jaenal, F.-A. Moreno, and J. Gonzalez-Jimenez, "Appearance-based sequential robot localization using a patchwise approximation of a descriptor manifold," *Sensors*, vol. 21, no. 7, p. 2483, 2021.

[29] T. K. Moon, "The Expectation-Maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[30] J.-L. Blanco, "A tutorial on SE(3) transformation parameterizations and on-manifold optimization," *University of Malaga, Tech.Rep.*, 2010.

[31] D. Fernandez-Chaves, J. Ruiz-Sarmiento, A. Jaenal, N. Petkov, and J. Gonzalez-Jimenez, "Robot@virtualhome, an ecosystem of virtual environment tools for realistic indoor robotic simulation," *Expert Systems with Applications*, 2022, submitted.

[32] T. Cieslewski, S. Choudhary, and D. Scaramuzza, "Data-efficient decentralized visual slam," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 2466–2473.

[33] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, 2019.

[34] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[35] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979.

[36] B. Wang, W. Shi, and Z. Miao, "Confidence analysis of standard deviational ellipse and its extension into higher dimensional euclidean space," *PloS one*, vol. 10, no. 3, p. e0118537, 2015.

[37] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" *arXiv preprint arXiv:2103.06443*, 2021.