Intelligent Reference Curation for Visual Place Recognition via Bayesian Selective Fusion

Timothy L. Molloy[1], Tobias Fischer[2], Michael Milford[2] and Girish N. Nair[1]

***Abstract*—A key challenge in visual place recognition (VPR) is recognizing places despite drastic visual appearance changes due to factors such as time of day, season, weather or lighting conditions. Numerous approaches based on deep-learnt image descriptors, sequence matching, domain translation, and probabilistic localization have had success in addressing this challenge, but most rely on the availability of carefully curated representative reference images of the possible places. In this paper, we propose a novel approach, dubbed Bayesian Selective Fusion, for actively selecting and fusing informative reference images to determine the best place match for a given query image. The selective element of our approach avoids the counterproductive fusion of every reference image and enables the dynamic selection of informative reference images in environments with changing visual conditions (such as indoors with flickering lights, outdoors during sunshowers or over the day-night cycle). The probabilistic element of our approach provides a means of fusing multiple reference images that accounts for their varying uncertainty via a novel training-free likelihood function for VPR. On difficult query images from two benchmark datasets, we demonstrate that our approach matches and exceeds the performance of several alternative fusion approaches along with state-of-the-art techniques that are provided with prior (unfair) knowledge of the best reference images. Our approach is well suited for long-term robot autonomy where dynamic visual environments are commonplace since it is training-free, descriptor-agnostic, and complements existing techniques such as sequence matching.**

***Index Terms*—Localization, Probabilistic Inference, Recognition, Autonomous Vehicle Navigation**

I. Introduction

VISUAL place recognition (VPR) is the problem of determining the place at which a given query image was captured and is a key enabling technology for mobile robot localization. The overwhelming majority of VPR approaches involve comparing query images with reference images previously captured at each candidate place in a map or database of the environment [1]. Research on VPR over several decades [1] has therefore explored techniques for robust image comparison including deep-learnt image descriptors [2]–[4], sequence matching [5]–[7], and multiprocess fusion [8].

The need for representative place images in VPR has also led to the development of techniques such as voting [9], experience maps [10], [11], memory compression [12] and domain translation [13] to curate references images. The majority of these approaches are learning-based, and determine the utility of reference images during training to select which to store or learn from. Determining the utility of reference images is however inherently challenging during training since ultimately it depends on factors affecting the similarity of the place appearance at query time. Surprisingly few approaches have sought to address this challenge since the rise of deep-

Fig. 1. Given a query image, Bayesian Selective Fusion selects and fuses informative reference images to find the best place match. Left: Given sunny query images, our approach selects images from the overcast and rain reference sets (frequently discarding the dusk images), and fuses them according to their likelihood (higher transparency indicates lower likelihood; see also black bars). Right: Given night query images, our approach selects images from the dusk reference set (frequently discarding the overcast and rain images) since they are perceptually most similar to night images. Note however that when dusk images offer limited visual information (e.g., due to sun glare), additional images from the rain and overcast references are selected and fused.

Manuscript received: October 15, 2020; Accepted December 13, 2020.

This paper was recommended for publication by Editor J. Civera upon evaluation of the Associate Editor and Reviewers' comments. This work received funding from the Australian Government, via grant AUSMURIB000001 associated with ONR MURI grant N00014-19-1-2571. T.F. and M.M. acknowledge continued support from the Queensland University of Technology (QUT) through the Centre for Robotics. *(Corresponding author: Timothy L. Molloy)*

[1]Timothy L. Molloy and Girish N. Nair are with the Department of Electronic and Electrical Engineering, University of Melbourne, Parkville, VIC 3010, Australia. {tim.molloy, gnair}@unimelb.edu.au.

[2]Tobias Fischer and Michael Milford are with the QUT Centre for Robotics, Queensland University of Technology, Brisbane, QLD 4000, Australia. {tobias.fischer, michael.milford}@qut.edu.au.

Digital Object Identifier (DOI): see top of this page.

learnt image descriptors for VPR in, e.g., [2], [7], [14]–[17].

Most recently, [18] has achieved state-of-the-art performance in a challenging VPR scenario with event-cameras using a training-free fusion approach for deep-learnt image descriptors. Although achieving state-of-the-art performance, the approach of [18] fuses all reference images at a given place. This complete fusion can be unnecessary and even counterproductive, especially in dynamic visual environments where query images can originate from strikingly different conditions — for example, indoors with lights flickering or outdoors with cloud and rain showers occluding the sun. In this paper, we propose a novel approach to intelligently fuse informative reference images to avoid unnecessary or counterproductive fusion.

The key contributions of this paper are:

1) A novel Bayesian Selective Fusion approach for *single-image* VPR that intelligently selects informative reference images and fuses these using Bayesian data fusion.
2) A novel training-free likelihood for probabilistic VPR that is agnostic to the underlying descriptors.
3) Demonstration of the state-of-the-art performance of our approach compared to the state-of-the-art approach of [18] and a suite of other fusion and non-fusion approaches.

Our approach complements existing VPR techniques (e.g. sequence matching) by providing a dynamic, training-free, descriptor-agnostic, and probabilistic means of exploiting the information from multiple (same-place) reference images.

This paper is structured as follows: Section II presents related work; Section III describes our proposed Bayesian Selective Fusion approach; Section IV reports our experimental setup; Section V presents the results of our experiments; and, Section VI offers conclusions and suggestions for future work.

II. Related Work

We first review probabilistic approaches to VPR and approaches that adapt the visual conditions of images before place matching. We then provide an overview of how previous approaches fuse multiple sets of reference images.

A. Probabilistic and Image Adaption Approaches

Arguably the most prominent probabilistic approach to recognizing places is FAB-MAP [19]. One of FAB-MAP's main innovations was to explicitly account for perceptual aliasing, such that highly similar but indistinct observations receive a low probability as being the same place; a generative model of bag-of-words observations implemented this. FAB-MAP has been extended in multiple ways, including the incorporation of odometry information and image sequences [20], [21].

While FAB-MAP relies on local image features, Lowry et al. [22] present a probabilistic model for whole-image descriptors, which leads to greater robustness in environments with large perceptual changes. Ramos et al. [23] propose a Bayesian framework for place recognition that leverages 5-9 training images per place and can then recognize a place from previously unseen viewpoints. Recently, Oyebode et al. [24] have proposed a Bayesian framework that leverages pre-trained object detectors to recognize indoor places based purely on object categories, which is related to the concept of visual place *categorization* [25], i.e. predicting the semantic category of a place.

Dubios et al. [26] use Bayesian filtering to model the dependency between sequences of images. Similar probabilistic temporal sequence modeling has also been investigated in [6] and in [27] using a Monte Carlo-based algorithm that achieves state-of-the-art performance on several datasets. In contrast to these methods that use sequences to temporally fuse place information, our VPR approach operates on single query images and fuses multiple reference images.

Our work is also related to approaches that preprocess query images to match conditions in reference images (or *vice versa*). For example, Pepperell et al. [28] and recent domain translation methods (cf. [13]) cast query images into the same visual conditions as the reference images via sky removal or night-to-day image translations (see also [29], [1, Section VII-A] and references therein). However, these approaches still raise the question of which reference images and translations are best to use for a given query image. For example, night-to-day translation implicitly assumes that query images are known to have been captured at night, and is unnecessary and could degrade performance on daytime query images.

B. Fusion of Reference Images

Kosecka et al. [9] described indoor places with a set of *representative views* and then found the most likely reference image using a maximum voting scheme. Carlevaris-Bianco and Eustice [30] proposed an extension to that approach by learning temporal observability relationships between the representative views. In a similar vein, the approach by Johns and Yang [31] learns feature co-occurrence statistics, such that each place is described using a set of co-occurring features. This allows for reliable matching of images that were captured at different times of the day.

When robots are deployed for prolonged periods, it is common to update the environment model over time. Biber and Duckett represent the environment using multiple timescales simultaneously, whereby older memories are faded at a rate that depends on the particular timescale being used [32]. For each place, the timescale with the highest log-likelihood is chosen. A more sophisticated approach in robot navigation is to store only persistent configurations in a long-term memory, so that dynamic objects do not become part of the map [33]. Another approach is the use of *coresets* [12] to drastically compress the images that represent a place. These coresets have subsequently been used to detect loop closures in real-time [34].

Impressive results have been achieved using so-called experience maps [10]. The accumulation of multiple experiences for the same place allows for both adaptation to changing appearance due to lighting and weather conditions as well as structural changes. At localization time, the robot can then predict the most appropriate experience using a probabilistic framework. Selecting the most appropriate experience results not only in reduced computational time but also in significantly reduced failure rates. A similar method by Doan et al. [11] accumulates additional experiences for the same place while avoiding an unbounded growth of computation and storage.

Vysotska and Stachniss [7] match places in a data association graph that can contain images of multiple reference traverses.

Further references to methods that contain multiple representations of the environment can be found in the survey on visual place recognition by Lowry et al. [1] in Section VII-B.2. While our method aims to leverage the complementary nature of multiple reference traverses, the multiprocess fusion approach by Hausler et al. [8] fuses predictions of multiple complementary image processing methods applied to the same input image. This also relates to [35], where appearance changes between the query and reference traversals are removed to some degree by using as few as 100 training samples. Contrary to [35], our method is training free. Our method is also related to [17], where in-sequence condition changes are investigated; however, only a single reference set is used.

III. Approach

In this section, we present our novel Bayesian Selective Fusion approach. We first revisit the formulation and solution of *single-image* VPR based on descriptor distances, including the state-of-the-art fusion approach of [18]. We exploit this formulation and the properties of descriptor distances to propose a novel strategy for selecting reference images and a new Bayesian method for fusing them, along with a novel likelihood function for probabilistic VPR.

A. Problem Formulation and the Minimum-Value Principle

Let X be an unknown place at which a robot captures a *query* image I_X and extracts a corresponding image descriptor vector $z_X \in \mathbb{R}^{N_z}$ with dimension N_z (e.g., using NetVLAD [2]). Let X belong to the set $\mathcal{X} = \{1, \dots, N\}$ of possible places. The robot has access to *reference* images of the places in $\mathcal{X}$ captured during previous visits under different visual appearance conditions. These reference images are stored in a total of M *reference sets*, with individual reference sets indexed by the scalar $u \in \{1, \dots, M\} \triangleq \mathcal{U}$. Each reference set $u \in \mathcal{U}$ contains (at most) one image from each place in $\mathcal{X}$ (together with associated image descriptor). Let I_i^u and $z_i^u \in \mathbb{R}^{N_z}$ denote the image and corresponding descriptor from reference set $u \in \mathcal{U}$ corresponding to place $i \in \mathcal{X}$. Given a reference set $u \in \mathcal{U}$, the robot can compute non-negative distances $d : \mathbb{R}^{N_z} \times \mathbb{R}^{N_z} \mapsto \mathbb{R}_{\geq 0}$ between the descriptor z_X of the query image I_X and the descriptors z_i^u from the reference set u for each place i in $\mathcal{X}$. We collect the distances corresponding to reference set u in the vector

$$D^u \triangleq [d(z_X, z_1^u) \quad d(z_X, z_2^u) \quad \cdots \quad d(z_X, z_N^u)]. \tag{1}$$

We shall use D_i^u to denote the ith component of D^u, i.e. $D_i^u \equiv d(z_X, z_i^u)$, and without loss of generality, we consider the Euclidean distance for $d(z_X, z_i^u)$ (other distance measures could also be used). The VPR problem is to infer the place X using the vectors D^u from some or all of the reference sets $\mathcal{U}$.

The vast majority of recent VPR approaches rely on the *minimum-value principle* (or best-match-strategy) [6], [22]. The principle asserts that the best place match $\hat{X}$ corresponds to the minimum descriptor distance in each reference set, i.e.,

$$X \approx \hat{X} = \arg\min_{i \in \mathcal{X}} D_i^u \tag{2}$$

for any $u \in \mathcal{U}$. The *minimum ensemble distance* approach of [18] is a recently proposed state-of-the-art method based on the minimum-value principle that fuses information from multiple reference sets. In this minimum ensemble distance approach, place matches $\hat{X}$ are found by minimizing the average (equivalently the total) of the descriptor distance vectors D^u over all reference sets, namely,

$$\hat{X} = \arg\min_{i \in \mathcal{X}} \sum_{u=1}^{M} D_i^u. \tag{3}$$

Clearly, VPR approaches based on the minimum-value principle perform poorest when the place that minimizes the descriptor distances is not the true place — a situation which for example might occur when the query and reference images are captured under different appearance conditions. Whilst their robustness can be improved by fusing reference sets captured under a variety of appearance conditions using addition or averaging (as in Eq. (3)), these operations implicitly assume that all reference sets provide equally useful information.

We hypothesize that reference sets will have different utility for different query images and that this utility will be inherently uncertain. Therefore, we propose a novel Bayesian selective fusion approach that: 1) selects informative reference sets based on the minimum-value principle; and, 2) fuses the selected reference sets using Bayesian fusion to handle the uncertainty of the minimum-value principle holding.

B. Proposed Reference Set Selection

Given the distance vectors D^u from all reference sets $u \in \mathcal{U}$, we select a variable subset of reference sets, defined as

$$\mathcal{S} = \left\{ u \in \mathcal{U} : \frac{\min_{i \in \mathcal{X}} D_i^u - \min_{i \in \mathcal{X}} D_i^{u^*}}{\min_{i \in \mathcal{X}} D_i^{u^*}} \leq \gamma \right\} \subset \mathcal{U}, \tag{4}$$

with which to compute a place decision. The collection $\mathcal{S}$ is formed by first selecting the "best" reference set under the minimum-value principle, that is, the reference set u^* that contains the minimum distance such that

$$u^* \triangleq \arg\min_{u \in \mathcal{U}} \min_{i \in \mathcal{X}} D_i^u. \tag{5}$$

All reference sets $u \in \mathcal{U}$ with minimum descriptor distances $\min_{i \in \mathcal{X}} D_i^u$ within a fraction $\gamma > 0$ of the minimum descriptors distance $\min_{i \in \mathcal{X}} D_i^{u^*}$ of the reference set u^* are also added to $\mathcal{S}$. Our reference set selection approach resembles outlier rejection with the parameter γ controlling the rejection of reference sets from $\mathcal{U}$ with minimum descriptor distances that are outliers compared to the "best" reference set u^*.

C. Proposed Bayesian Fusion

Given the descriptor distance vectors $\mathcal{D}^{\mathcal{S}} \triangleq \{D^u : u \in \mathcal{S}\}$ from the selected reference sets $\mathcal{S}$, the Bayesian belief $P(X|\mathcal{D}^{\mathcal{S}})$ over the places in $\mathcal{X}$ is given by Bayes' rule:

$$P(X = i|\mathcal{D}^{\mathcal{S}}) = \frac{P(\mathcal{D}^{\mathcal{S}}|X = i)P(X = i)}{\sum_{j=1}^{N} P(\mathcal{D}^{\mathcal{S}}|X = j)P(X = j)} \tag{6}$$

for $i \in \mathcal{X}$. Any prior knowledge that the robot has about its location (arising for example from its motion model or

previous place matches) is incorporated here through the prior probability distribution $P(X)$. In the absence of any prior knowledge, this is taken as uniform (i.e. $P(X=i)=N^{-1}$ for all $i \in \mathcal{X}$). The likelihood $P(\mathcal{D}^{\mathcal{S}}|X)$ enables our novel fusion of the place information from the selected references $\mathcal{S}$.

To make the novel fusion operation in Eq. (6) explicit (and scalable), note that the descriptor distances D_i^u are solely functions of the reference image descriptors z_i^u after X (and hence I_X and z_X) is given. The vectors D^u from different reference sets are thus conditionally independent given X, and the likelihood $P(\mathcal{D}^{\mathcal{S}}|X)$ simplifies to the product:

$$P(\mathcal{D}^{\mathcal{S}}|X=i) = \prod_{u\in\mathcal{S}} P(D^u|X=i) \tag{7}$$

for $i \in \mathcal{X}$ where $P(D^u|X)$ are the likelihoods for the single (individual) reference sets $u \in \mathcal{S}$. Constructing the *single-reference* likelihoods $P(D^u|X)$ is easier and more scalable than constructing the joint likelihood $P(\mathcal{D}^{\mathcal{S}}|X)$ as we shall discuss in the following subsection. Here, we note that since the likelihood $P(\mathcal{D}^{\mathcal{S}}|X)$ in Eq. (7) has a product form and the denominator in Eq. (6) serves only to normalize, we may rewrite Eq. (6) as simply:

$$P(X=i|\mathcal{D}^{\mathcal{S}}) \propto \prod_{u\in\mathcal{S}} P(D^u|X=i)P(X=i) \tag{8}$$

for $i \in \mathcal{X}$. The robot's place recognition decision with our approach, denoted $\hat{X}$, is then the place with the maximum (unnormalized) Bayesian belief, namely,

$$\hat{X} = \arg\max_{i\in\mathcal{X}} \prod_{u\in\mathcal{S}} P(D^u|X=i)P(X=i). \tag{9}$$

No place recognition decision is made if the belief fails to exceed a threshold $h > 0$, which balances recognizing the wrong place (small h) with missing the true place (large h). We next describe a novel efficient construction of the *single-reference* likelihoods $P(D^u|X)$ used in Eq. (9).

D. Proposed Training-Free Single-Reference Likelihoods

To construct the *single-reference* likelihoods $P(D^u|X)$, note that by definition they are the joint likelihoods of the distances, i.e., $P(D^u|X) \equiv P(D_1^u, \ldots, D_N^u|X)$. Note also that in order for the distances D_i^u to all be equally useful, their (marginal) distribution should depend only on whether the places X and i match or not, and not on the specific places X and i. Hence, we model the distances D_i^u as conditionally independent given X and distributed according to a (marginal) distribution $P(D_i^u|X \neq i)$ when the place X does not match i and a different (marginal) distribution $P(D_i^u|X=i)$ when the place X matches i. The (joint) likelihood of any place i after observing the vector D^u from reference set u is thus the product of the (marginal) likelihoods after observing the individual distances, namely,

$$P(D^u|X=i) = P(D_i^u|X=i) \prod_{j=1, j\neq i}^{N} P(D_j^u|X\neq j)$$

for $i \in \mathcal{X}$. Equivalently, we may write this likelihood as

$$P(D^u|X=i) \propto \frac{P(D_i^u|X=i)}{P(D_i^u|X\neq i)} \tag{10}$$

for $i \in \mathcal{X}$. Here, the proportionality constant $C^u \triangleq \prod_{j=1}^{N} P(D_j^u|X\neq j)$ is constant for all $i \in \mathcal{X}$, which allows us to simply ignore it in our approach (cf. Eq. (9)).

Place-match and non-place-match likelihoods, $P(D_i^u|X=i)$ and $P(D_i^u|X\neq i)$, have previously been constructed via training for (whole-image) descriptors (cf. [1], [22]). These approaches can be used directly to evaluate Eq. (10), however we propose an alternative training-free approach based on a probabilistic encoding of the minimum-value principle (Eq. (2)). For a given reference set u, we model the place-match likelihood $P(D_i^u|X=i)$ as being proportional to the number of places with descriptor distances larger than D_i^u in D^u, i.e.,

$$P(D_i^u|X=i) \propto \sum_{j=1}^{N} \mathbb{1}\{D_i^u \leq D_j^u\} \tag{11}$$

where the indicator function $\mathbb{1}\{D_i^u \leq D_j^u\}$ takes a value of 1 when $D_i^u \leq D_j^u$ and is zero otherwise. Again, the proportionally constant is independent of i, and can be ignored.

We model the non-place-match likelihood $P(D_i^u|X\neq i)$ by recalling that the distribution of descriptor distances is independent of the place, and hence most elements in the vector D^u are realizations of the descriptor distance with $X \neq i$. Due to the computational efficiency of computing the mean μ and variance σ^2 of the elements in D^u, we simply model the non-place-match likelihood as the Gaussian $P(D_i^u|X\neq i) = \mathcal{N}(D_i^u; \mu, \sigma^2)$ (accepting that some bias will be incurred due to one element i of D^u corresponding to $X=i$).

E. Summary of Bayesian Selective Fusion

In summary, our proposed approach first selects reference sets according to Eq. (4), then computes the single-reference likelihoods given by Eq. (10) using Eq. (11) and our Gaussian construction of non-place-match likelihood, and finally declares a place recognition decision via the maximization in Eq. (9).

IV. Experimental Setup

In this section, we describe the datasets, methods, metrics, and parameters with which we evaluate our approach.

A. Datasets

We use two, widely used, established benchmark datasets.

The **Nordland** dataset [36] consists of four recordings of a train traversing a 728km long route in Norway captured in spring, summer, fall and winter. We extract one image per second from a one hour and twenty minute long section of these videos (20:00 to 1:40:00) and, as typically done in the literature, manually remove tunnels and sections where the train is stationary or traveling at speeds below 15km/h (based on the available GPS information). This resulted in a total of 3000 frames per season. As frame i in one of those videos corresponds exactly to frame i in the other videos, we find ground-truth correspondences by matching the query traverse frame number with the reference traverse number and allow a ground-truth tolerance of ± 2 frames. The Nordland dataset has been widely used in the literature, e.g. in [4], [8], [35].

The **Oxford RobotCar** dataset [37] contains over 100 traversals of a consistent route through Oxford, captured at different times of the day and in varying weather conditions. Recently, centimeter-accurate ground-truth annotations (based on post-processed GPS, IMU and GNSS base station recordings) were made available for a subset of these traversals [38]. We selected five traversals of the same route[1] and subsampled them such that places i and $i+1$ have a regular spatial separation of one meter between them, which resulted in 3350 frames. We used images recorded using the front left stereo camera. Similarly to the Nordland dataset, the Oxford RobotCar dataset is widely used in the literature, e.g. see [4], [8]. As in [4], we use a ground-truth tolerance of ± 10 meters.

We computed NetVLAD [2] descriptors for the images extracted from the datasets. Following [2] and as has become standard in subsequent studies (cf. [4], [18]), the descriptors were reduced to 4096 components via PCA. We use these NetVLAD descriptors for z_X and z_i^u in all experiments except for one comparative study (Section V-F), for which we also compute 4096-d DenseVLAD [3] descriptors.

B. Baseline, Fusion, and Single-Reference Methods

We use the state-of-the-art minimum ensemble distance approach of [18] (see also Eq. (3)) as the *Baseline Fusion* approach for comparison with our proposed *Bayesian Selective Fusion* approach. The reference sets used by these fusion methods constitute the remainder of the traversals on the relevant dataset after removing the query traversal (i.e., if the query is Nordland Summer, then the reference sets are Winter, Fall and Spring). We also implement standard *single-reference* NetVLAD (cf. Eq. (2)) and versions of our Bayesian approach with single reference sets (i.e. $M = 1$). These single-reference methods are labeled according to their reference set (e.g., given only a Nordland spring reference, NetVLAD is termed *NetVLAD Spring* and the single-reference Bayesian approach is termed *Bayesian Spring* or simply *Spring*).

C. Performance Metrics

We report VPR performance using Precision-Recall (PR) curves (with the threshold h swept for our proposed method). Precision is defined as the ratio of correct place matches to the total number of place matches whilst recall is defined as the ratio of correct place matches to the total number of possible true matches. Our datasets have a true match for every query image. In some experiments, we also report the area under the PR curve (AUC) as a summary statistic. As in [14]–[16], the AUC serves as a proxy for recall at 100% precision (which is of direct concern for SLAM) for comparing single-image VPR methods since many offer no recall at 100% precision.

D. Parameter: Fraction for Selecting Reference Sets γ

Our approach relies on a single parameter, γ (see Eq. (4)). In our experience, values of $\gamma \in [0.04, 0.1]$ provide reasonable performance on the Nordland and Oxford datasets. To illustrate this insensitivity, we use a fixed $\gamma = 0.04$ in all experiments, noting that improvements could be obtained by tuning γ.

[1]Dusk: 2014-11-21-16-07-03, Night: 2014-12-16-18-44-24, Overcast: 2015-05-19-14-06-38, Sun: 2015-08-12-15-04-18, Rain: 2015-10-29-12-18-17

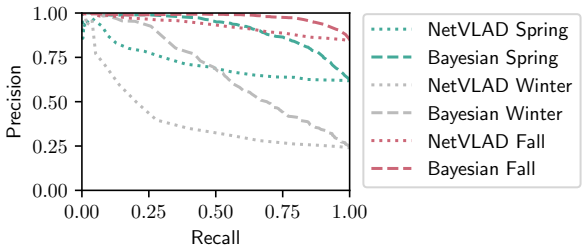


Fig. 2. Precision-recall of single-reference (no fusion) Baseline and proposed Bayesian methods on a Nordland Summer query traversal.

V. Experimental Results

In this section, we evaluate our approach and its reference selection and Bayesian aspects across six experiments, including comparisons and extensions to state-of-the-art techniques. In Section V-A, we first examine the performance offered by our new training-free likelihood for Bayesian VPR in comparison to NetVLAD. In Sections V-B and V-C, we compare our Bayesian Selective Fusion approach with the state-of-the-art fusion approach of [18] and NetVLAD (with NetVLAD provided advantageously with the best performing reference set). In Section V-D we showcase our approach in a scenario with extreme appearance changes. Finally, in Sections V-E and V-F we demonstrate that our approach can also exploit sequence matching (i.e. SeqSLAM [5]) and alternative image descriptors (i.e. DenseVLAD [3]).

A. Single-Reference Method Comparison

Fig. 2 shows that Bayesian approaches with our new training-free *single-reference* likelihood outperforms NetVLAD given the same reference set for query images drawn from the Nordland summer traversal. The Bayesian approach offers a 5.5% (relative) improvement in AUC over NetVLAD when both perform their best (Fall reference) and a 70% improvement when both perform their worst (Winter reference). Fig. 2 motivates the use of (selective) fusion since the performance of the methods is strongly dependent on the reference images they are provided. We thus now focus on evaluating the performance of our Bayesian Selective Fusion approach.

B. Bayesian Selective Fusion vs Baseline Fusion

We next compare our Bayesian Selective Fusion approach with the state-of-the-art fusion approach of [18] (i.e., Baseline Fusion) and NetVLAD (unfairly) given the best performing reference on each traversal. Figs. 3 and 4 show that our approach outperforms the Baseline approach by a large margin on Nordland Summer and Winter query traversals with AUCs of 0.97 and 0.84 for our approach compared to 0.89 and 0.68 for the baseline method (9% and 24% increase, respectively). Our approach performs slightly worse than the Baseline Fusion approach on the easier Oxford Sun traversal (with an AUC of 0.995 compared to 0.997) but outperforms it on the more

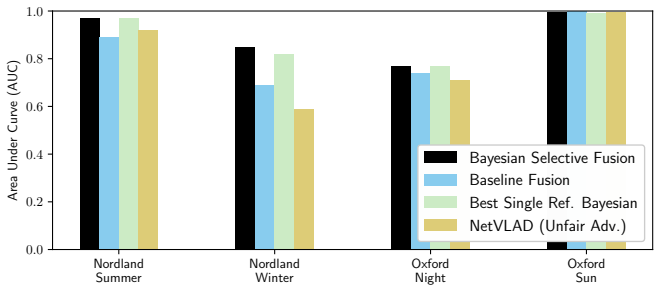


Fig. 3. Area Under Precision-Recall Curves of Fig. 4 for query images from four traverses from Nordland and Oxford RobotCar datasets.

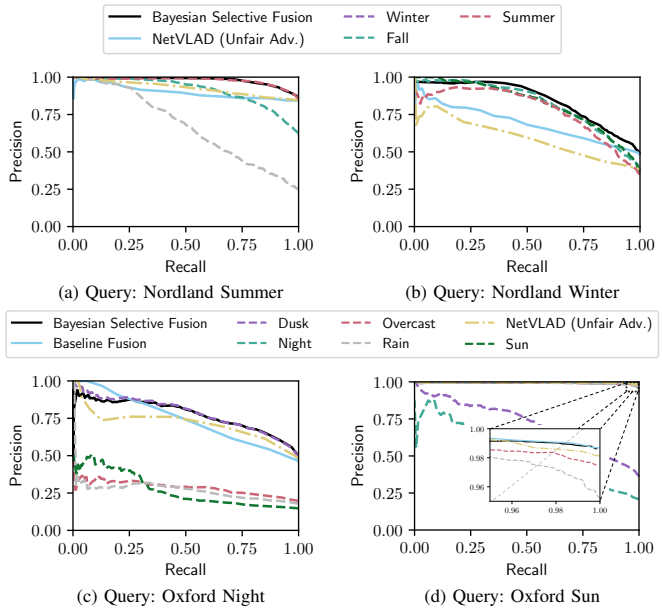


Fig. 4. Precision-recall comparison of Bayesian Selective Fusion, Baseline Fusion, and single-reference Bayesian methods on query images from four traverses from Nordland and Oxford RobotCar datasets.

difficult Oxford Night traversal, with an AUC of 0.77 compared to 0.74 (4% increase).

Our approach can be seen to select informative reference sets (also shown in Fig. 1) since it attains equal or better performance than the *single-reference* Bayesian methods. Such a scenario is shown in Fig. 4(b) on Nordland Winter, where our method's AUC is higher than the mean AUC of the single-reference Bayesian methods by 5% and the best by 3%. Finally, our approach outperforms NetVLAD on all but the easiest case (i.e., Oxford Sun with an absolute difference in AUC of less than 0.003); improvements in AUC range from 8.5% on Oxford Night to 25% on the more difficult Nordland Winter. Examples of correct matches from our approach and false matches from NetVLAD are shown in Fig. 10.

C. Comparison with Alternative Fusion Approaches

We now perform an ablation-type study to examine the performance of our approach with and without Bayesian fusion and/or reference set selection. We consider a *Bayesian Full Fusion* approach that omits reference selection; and, a *Baseline Selective Fusion* approach based on the Baseline Fusion method

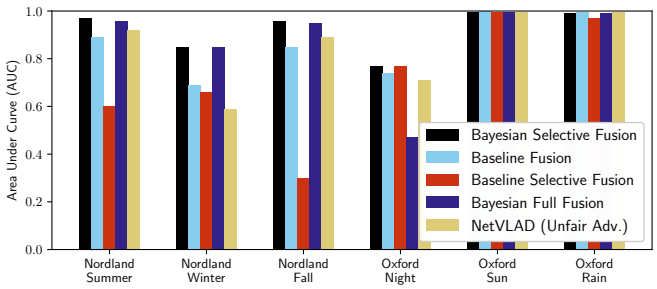


Fig. 5. Area Under Precision-Recall Curves of Fig. 6 for query images from six traversals from Nordland and Oxford RobotCar datasets.

of Eq. (3) that uses only the selected reference sets $\mathcal{S}$. We also consider NetVLAD (advantageously) given the best reference on each traversal.

Figs. 5 and 6 show that only our proposed Bayesian Selective Fusion approach works consistently well without any modification or parameter tuning across six query traversals from the Nordland and Oxford datasets. All other approaches have instances of total failure or inferior performance. For example, the Baseline Selective Fusion method fails almost completely on the difficult Nordland Summer, Winter, and Fall traversals, whilst the Baseline Fusion method and NetVLAD perform poorly. The Bayesian Full Fusion approach compares the most favorably with our proposed Bayesian Selective Fusion approach but does not outperform it, illustrating the additional benefit of reference set selection.

D. Extreme Appearance Changes

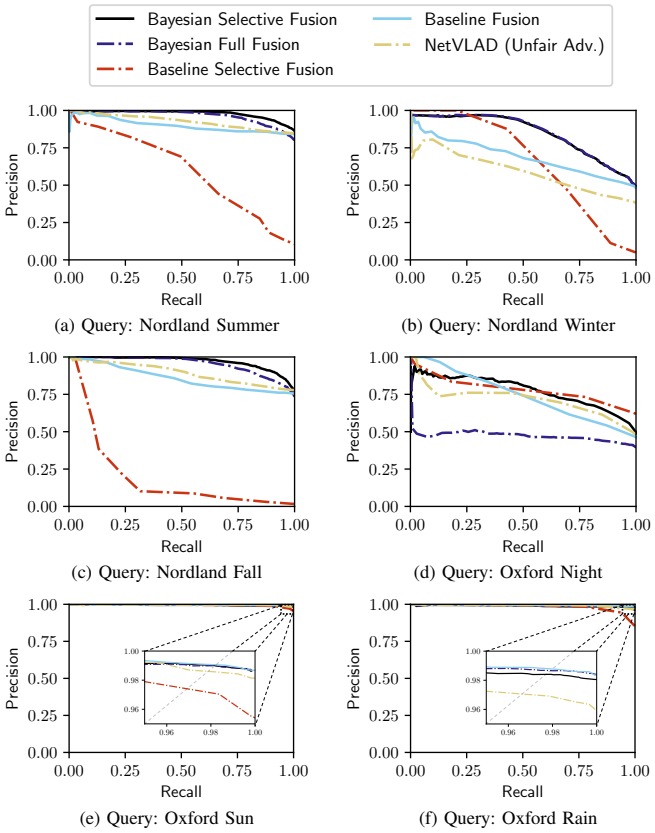
Our reference set selection approach is motivated in part by the need for VPR systems to adapt in situations where the visual appearance of places can change dynamically and unpredictably such as indoors with a flickering light. For this study, we consider a scenario in which the query image at each place in the Oxford dataset is drawn randomly (with equal probability) from either the Sun or Night traversals.

The performance of our proposed approach on the randomized Sun/Night query traversal is shown in Fig. 7 together with the performance of the Dusk and Overcast single-reference Bayesian methods that perform best on the standalone Oxford Sun and Night traversals. We also consider NetVLAD with it again given the advantage of the best reference set (overcast). From Fig. 7, we see that Bayesian Selective Fusion outperforms the single-reference approaches and (the advantaged) NetVLAD with higher precision at recalls greater than 0.9, and AUC improvements over (the advantaged) NetVLAD, Overcast, and Dusk of 1.0%, 6.6%, and 29%, respectively.[2]

E. Sequence Matching

Sequence matching techniques are commonly used in conjunction with single-image place matching for VPR under

[2]We expect worse performance from single-reference methods when query images can originate from more than two conditions (e.g. Raining Night/Day, Snowing Night/Day, Raining Dusk/Dawn, etc.) but we lack sufficient data to perform these studies.



(a) Query: Nordland Summer (b) Query: Nordland Winter

(c) Query: Nordland Fall (d) Query: Oxford Night

(e) Query: Oxford Sun (f) Query: Oxford Rain

Fig. 6. Precision-recall comparison of Bayesian Selective Fusion approach with alternative fusion approaches on query images from six traversals from Nordland and Oxford RobotCar datasets.

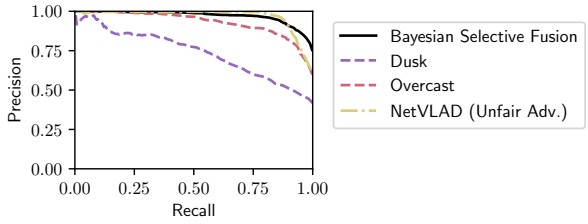


Fig. 7. Precision-recall on randomized query images from Sun and Night traversals from the Oxford RobotCar dataset.

challenging environments. Fig. 8 shows that the performance of our approach on the randomized query from our previous study is improved via integration with an existing sequence matching approach (SeqSLAM [5]) with a sequence length of 5. It also performs better than NetVLAD integrated with the same sequence matching approach despite NetVLAD having prior knowledge of the best reference set.

F. Other Image Descriptors

In our previous studies, we used NetVLAD [2] descriptors. Fig. 9 shows the results of repeating our study from Section V-B for Nordland summer and winter traversal queries with DenseVLAD [3] descriptors instead of NetVLAD descriptors. We see that our approach delivers a similar performance gain compared to the best single-reference Bayesian methods without any modification, parameter tuning, or training.

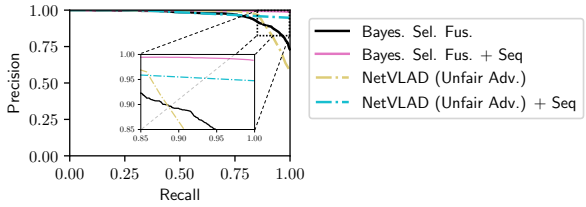
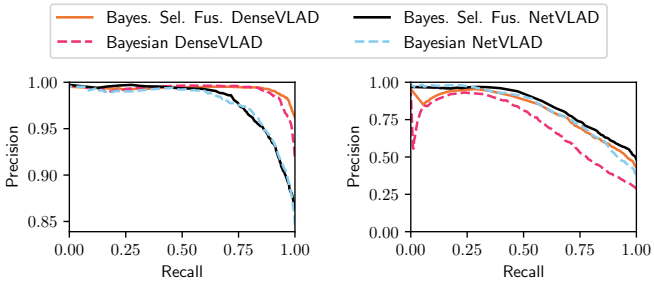


Fig. 8. Precision-recall of methods with and without sequence matching on randomized query images from Sun and Night traversals from the Oxford RobotCar dataset.



(a) Nordland Summer (b) Nordland Winter

Fig. 9. Precision-recall of Bayesian Selective Fusion and best single-reference Bayesian methods on query images from Nordland Summer and Winter traversals using DenseVLAD or NetVLAD image descriptors.

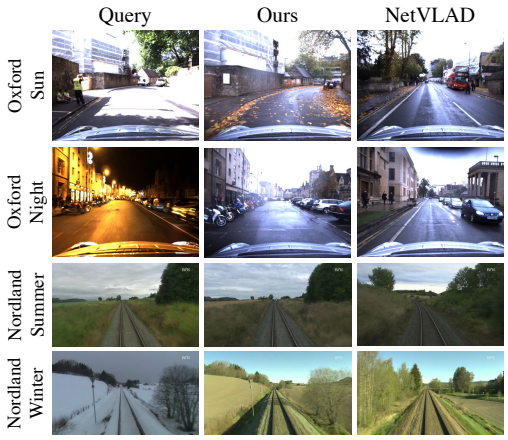


Fig. 10. Qualitative examples where our Bayesian Selective Fusion approach successfully localizes, while NetVLAD produces incorrect place matches, even when given the best reference set.

G. Compute Time and Scaling

On an i7-8750H CPU, our approach took an average of 0.0224s (0.0002s for reference selection and 0.0222s for fusion) per query image with $M = 3$ and $N = 3000$ places. Complexity is linear in the number of references M, however, the expensive Bayesian operations (Eq. (9) and Eq. (10)) scale linearly with the number of *selected* reference sets in $\mathcal{S}$.

VI. Conclusions

Visual place recognition relies on comparing query images with reference images previously captured and stored. The storage and exploitation of multiple same-place reference images has garnered surprisingly little recent attention, with state-of-the-art approaches either attempting to determine the utility of reference images *a priori* without consideration of

the conditions at query time, or fusing all available reference images, which can be unnecessary and counterproductive. In this paper, we have proposed Bayesian Selective Fusion as a novel approach to dynamically determine and exploit the utility of reference images at query time. We have demonstrated that our approach can exceed the performance of state-of-the-art techniques on challenging benchmark datasets, including state-of-the-art single-reference methods provided with (advantageous) prior knowledge of the best reference images. We have also shown that our approach is descriptor-agnostic and can be used in conjunction with standard techniques including sequence matching to achieve state-of-the-art performance.

Our current work can be extended in several ways, including the use of Bayesian Selective Fusion with recently proposed Delta Descriptors [4], local descriptors, and probabilistic sequence matching via the prior in Eq. (6). We are also interested in evaluating our approach in situations with large viewpoint variations between the query and reference images but are currently limited by existing datasets offering too few reference sets. Furthermore, it would be interesting to explore the insight from our reference set selection approach to determine which extra reference sets should be collected, or which can be discarded to minimize storage requirements. Finally, we believe that our research contributes to a better understanding of how to exploit the information from deep-learnt image descriptors within a training-free Bayesian setting, opening the possibility of using temporal Bayesian filtering in conjunction with reference set fusion, and further exploiting information-theoretic techniques in visual place recognition.

REFERENCES

[1] S. Lowry *et al.*, "Visual Place Recognition: A Survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, 2016.

[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, 2018.

[3] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 Place Recognition by View Synthesis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 257–271, 2018.

[4] S. Garg, B. Harwood, G. Anand, and M. Milford, "Delta descriptors: Change-based place representation for robust visual localization," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5120–5127, 2020.

[5] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE Int. Conf. Robot. Autom.*, 2012, pp. 1643–1649.

[6] T. Naseer, W. Burgard, and C. Stachniss, "Robust visual localization across seasons," *IEEE Trans. Robot.*, vol. 34, no. 2, pp. 289–302, 2018.

[7] O. Vysotska and C. Stachniss, "Effective visual place recognition using multi-sequence maps," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1730–1736, 2019.

[8] S. Hausler, A. Jacobson, and M. Milford, "Multi-process fusion: Visual place recognition using multiple image processing methods," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1924–1931, 2019.

[9] J. Košecká, F. Li, and X. Yang, "Global localization and relative positioning based on scale-invariant keypoints," *Robot. Auton. Syst.*, vol. 52, no. 1, pp. 27–38, 2005.

[10] C. Linegar, W. Churchill, and P. Newman, "Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation," in *IEEE Int. Conf. Robot. Autom.*, 2015, pp. 90–97.

[11] A.-D. Doan, Y. Latif, T.-J. Chin, Y. Liu, T.-T. Do, and I. Reid, "Scalable Place Recognition Under Appearance Change for Autonomous Driving," in *IEEE Int. Conf. Computer Vision*, 2019, pp. 9319–9328.

[12] G. Rosman, M. Volkov, D. Feldman, J. W. Fisher III, and D. Rus, "Coresets for k-segmentation of streaming data," in *Advances Neural Information Process. Syst.*, 2014, pp. 559–567.

[13] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. V. Gool, "Night-to-day image translation for retrieval-based localization," in *IEEE Int. Conf. Robot. Autom.*, 2019, pp. 5958–5964.

[14] Z. Chen *et al.*, "Deep learning features at scale for visual place recognition," in *IEEE Int. Conf. Robot. Autom.*, 2017, pp. 3223–3230.

[15] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight CNNs for significant viewpoint and appearance changes," *IEEE Trans. Robot.*, vol. 36, no. 2, pp. 561–569, 2019.

[16] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2017, pp. 9–16.

[17] S. Schubert, P. Neubert, and P. Protzel, "Unsupervised learning methods for visual place recognition in discretely and continuously changing environments," in *IEEE Int. Conf. Robot. Autom.*, 2020, pp. 4372–4378.

[18] T. Fischer and M. Milford, "Event-Based Visual Place Recognition With Ensembles of Temporal Windows," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6924–6931, 2020.

[19] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, 2008.

[20] W. Maddern, M. Milford, and G. Wyeth, "CAT-SLAM: probabilistic localisation and mapping using a continuous appearance-based trajectory," *Int. J. Robot. Res.*, vol. 31, no. 4, pp. 429–451, 2012.

[21] W. Maddern, M. Milford, and G. Wyeth, "Towards persistent localization and mapping with a continuous appearance-based topology," in *Robot.: Sci. Syst.*, 2012, pp. 281–288.

[22] S. M. Lowry, G. F. Wyeth, and M. J. Milford, "Towards training-free appearance-based localization: probabilistic models for whole-image descriptors," in *IEEE Int. Conf. Robot. Autom.*, 2014, pp. 711–717.

[23] F. Ramos, B. Upcroft, S. Kumar, and H. Durrant-Whyte, "A Bayesian approach for place recognition," *Robot. Auton. Syst.*, vol. 60, no. 4, pp. 487–497, 2012.

[24] K. Oyebode, S. Du, B. J. Van Wyk, and K. Djouani, "A sample-free bayesian-like model for indoor environment recognition," *IEEE Access*, vol. 7, pp. 79 783–79 790, 2019.

[25] J. Wu, H. I. Christensen, and J. M. Rehg, "Visual place categorization: Problem, dataset, and algorithm," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2009, pp. 4763–4770.

[26] M. Dubois, H. Guillaume, F. Emmanuelle, and P. Tarroux, "Visual place recognition using Bayesian filtering with Markov chains," in *Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, 2011, pp. 435–440.

[27] A.-D. Doan *et al.*, "Visual localization under appearance change: filtering approaches," *Neural Comput. Appl.*, 2020, to appear.

[28] E. Pepperell, P. Corke, and M. Milford, "Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints," *Int. J. Robot. Res.*, vol. 35, no. 9, pp. 1057–1179, 2016.

[29] O. Saurer, G. Baatz, K. Köser, M. Pollefeys *et al.*, "Image based geo-localization in the alps," *Int. J. Comput. Vision*, vol. 116, no. 3, pp. 213–225, 2016.

[30] N. Carlevaris-Bianco and R. M. Eustice, "Learning temporal coobservability relationships for lifelong robotic mapping," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. Workshops*, 2012.

[31] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *IEEE Int. Conf. Robot. Autom.*, 2013, pp. 3212–3218.

[32] P. Biber and T. Duckett, "Experimental analysis of sample-based maps for long-term slam," *Int. J. Robot. Res.*, vol. 28, no. 1, pp. 20–33, 2009.

[33] T. Morris, F. Dayoub, P. Corke, G. Wyeth, and B. Upcroft, "Multiple map hypotheses for planning and navigating in non-stationary environments," in *IEEE Int. Conf. Robot. Autom.*, 2014, pp. 2765–2770.

[34] M. Volkov, G. Rosman, D. Feldman, J. W. Fisher, and D. Rus, "Coresets for visual summarization with applications to loop closure," in *IEEE Int. Conf. Robot. Autom.*, 2015, pp. 3638–3645.

[35] S. Lowry and M. J. Milford, "Supervised and unsupervised linear learning techniques for visual place recognition in changing environments," *IEEE Trans. Robot.*, vol. 32, no. 3, pp. 600–613, 2016.

[36] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," in *IEEE Int. Conf. Robot. Autom. Workshops*, 2013.

[37] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2017.

[38] W. Maddern, G. Pascoe, M. Gadd, D. Barnes, B. Yeomans, and P. Newman, "Real-time kinematic ground truth for the Oxford RobotCar dataset," *arXiv preprint arXiv: 2002.10152*, 2020.