Delta Descriptors: Change-Based Place Representation for Robust Visual Localization

Sourav Garg\textsuperscript{1}, Ben Harwood\textsuperscript{2}, Gaurangi Anand\textsuperscript{3} and Michael Milford\textsuperscript{1}

Abstract—Visual place recognition is challenging because there are so many factors that can cause the appearance of a place to change, from day-night cycles to seasonal change to atmospheric conditions. In recent years a large range of approaches have been developed to address this challenge including deep-learnt image descriptors, domain translation, and sequential filtering, all with shortcomings including generality and velocity-sensitivity. In this paper we propose a novel descriptor derived from tracking changes in any learned global descriptor over time, dubbed \textit{Delta Descriptors}. Delta Descriptors mitigate the offsets induced in the original descriptor matching space in an unsupervised manner by considering temporal differences across places observed along a route. Like all other approaches, Delta Descriptors have a shortcoming - volatility on a frame to frame basis - which can be overcome by combining them with sequential filtering methods. Using two benchmark datasets, we first demonstrate the high performance of Delta Descriptors in isolation, before showing new state-of-the-art performance when combined with sequence-based matching. We also present results demonstrating the approach working with a second different underlying descriptor type, and two other beneficial properties of Delta Descriptors in comparison to existing techniques: their increased inherent robustness to variations in camera motion and a reduced rate of performance degradation as dimensional reduction is applied. Source code will be released upon publication.

I. INTRODUCTION

Visual Place Recognition (VPR) is one of the key enablers for mobile robot localization and navigation. The earlier approaches to VPR predominantly relied on hand-crafted local (SIFT [1], ORB [2]) and global (HoG [3], GIST [4]) image representation methods. The use of local features in BoW [5] and VLAD [6] like encoding techniques based on visual vocabularies has been a popular choice for the task of global place retrieval (kidnapped robot problem). However, the lack of appearance robustness of underlying local features led to the development of appearance-invariant whole-image description techniques, combined with sequence-based matching [7], [8] in order to deal with extreme appearance variations.

Recent advances in deep learning have led to more robust counterparts to hand-crafted local and global feature representations like LIFT [9], DeLF [10], NetVLAD [11], and LoST [12]. Furthermore, GANs [13] based night-to-day image translation [14], feature fusion [15], and teacher-student networks [16] have also been explored for VPR. Although achieving state-of-the-art performance, deep learning based image descriptors often suffer from the challenges of data bias that limits their utility for out-of-the-box operations. This is typically resolved by re-training or fine-tuning the CNN. However, this may not always be feasible for all the application scenarios like VPR: supervised learning would require multiple traverses of the new environment such that it captures variations in scene appearance and camera viewpoint.

In this paper, we propose an unsupervised method for transforming existing deep-learnt image descriptors into change-based representations, dubbed \textit{Delta Descriptors}. Due to appearance variations in the environment during a revisit, the original descriptors tend to be offset by a margin, leading to an increased distance between descriptors belonging to the same place. Defined in a difference space [18], Delta Descriptors implicitly deal with such offsets, resulting in a reduced distance between the same place descriptors. This is of particular importance to a mobile robot or autonomous vehicle operating in a new environment that may undergo significant variations in visual appearance during a repeated route traversal. Previously, difference-based approaches have been explored for recognizing objects [18], faces [19] and actions [20]. In this research, the concept of difference-based description is based on measuring the changes across an observed sequence of places which is repeatable across traverses.

Using the proposed Delta Descriptors, we:

- establish that given a fixed sequential span, Delta Descriptors perform at par with sequence-based matching, and
- that they achieve state-of-the-art performance when used...
in conjunction with the sequence-based matching,
  - show Delta Descriptors retain high recall performance when dimensional reduction techniques like PCA are used, in contrast to raw descriptors, especially in the presence of strong appearance variations,
  - demonstrate their robustness to variations in camera motion both along and across the repeated traverses, unlike sequence-based matching methods which either require motion information or a sophisticated sequence-search approach, and
  - provide insights into selecting sequential span sizes for calculating Delta Descriptors, the role of order in which places are observed, an investigative Multi-Delta Descriptor approach to deal with velocity variability, and image-level visualization of how the proposed descriptors aid in time-series pattern matching.

The paper is divided into the following sections: Section II discusses the prior literature and related work; Section III describes the proposed approach for calculating Delta Descriptors; Section IV details the experimental setup including dataset description, system parameter estimation, and evaluation methods; Section V presents the results on the benchmark datasets and characterizes the proposed descriptor through a range of experiments; Section VI dives into the visualization of CNN activations of image regions for Delta Descriptors; and Section VII concludes the paper, highlighting potential ways to extend the current work in future.

II. RELATED WORK

A. Hand-crafted and Deep-Learnt Descriptors

The ability to represent images as a compact descriptor remains a key requirement for VPR. Broadly speaking, the purpose of these descriptors is to map images of a particular physical location into a lower dimensional representation. In the context of VPR, the goal of these mappings is to preserve a unique description of each location while removing changing information such as camera orientation, ambient lighting, mobile distractors and seasonal variations.

In the era of predominantly hand-crafted descriptors, the ORB [2] descriptor was designed to utilise a training set of local image patches. These patches were used to generate a set of binary tests that would maximise rotational invariance. While this learning procedure enabled the creation of more computationally efficient features, it also introduced the usual biases that come from training a model on a particular data set.

With the advent of deep learning, hand-crafted descriptors such as SIFT [1] have largely been replaced by learned descriptors such as LIFT [9] and DeLF [10]. Learning these local patch-based representations end-to-end has again yielded improved descriptors, but has also increased the reliance on having training data that accurately models all aspects of the target domain. This data bias is also seen in learned global representations such as NetVLAD [11]. However, this problem becomes more challenging to address since it is more challenging to collect a representative data set of place images across all invariances than it is for image patches.

B. Sequence-based Representations

A vast literature exists for spatio-temporal representation of video data with applications in action classification [21], activity recognition [22], person re-identification [23], dense video captioning [24], 3D semantic labelling [25], 3D shape completion [26], and dynamic scene recognition [27]. This has led to the emergence of 3D CNNs [28], [29] involving 3D convolutions to learn better spatio-temporal representations that are suited to the task at hand. However, most of the aforementioned tasks only require dealing with a limited number of classes unlike VPR where every other place is a unique observation. Furthermore, methods based on RNNs, LSTM networks [30] and GRUs [31] tend to learn general patterns of how temporal information is ordered, for example, in the case of action or activity recognition. For VPR, such general patterns of order may not exist beyond the overlapping visual regions around a particular visual landmark.

In the context of VPR, there have been some attempts towards developing robust spatio-temporal or sequence-based representations. In [32], authors learnt spatio-temporal landmarks based on SURF features, but these local features needed to be independently tracked on frame-to-frame basis. In [33], authors proposed a bio-inspired place recognition method that used environment-specific discriminative training of different Long-Term Memory (LTM) cells. [34] explored three novel techniques: Descriptor Grouping, Descriptor Fusion, and Recurrent Descriptors, to accrue deep features from multiple views of the scene across the route. [35] proposed a topometric spatio-temporal representation of places using monocular depth estimation, mainly focused on recognizing places from opposing viewpoints. [36] proposed coresets-based visual summarization method for efficient hierarchical place recognition. Doing away with a compact representation, [37] used a variety of image descriptors to represent groups of images and formulated sequence-searching as an optimization problem.

C. Sequence-based Matching

Although sequence-based representations are not that common in VPR literature, use of sequence-based matching has been extensively explored for VPR. Such methods leverage sequential information after computing the place matching scores where place representations are typically based on a single image. This leads to enhanced VPR performance [38], particularly in cases where perceptual aliasing is very high, for example, dealing with extreme appearance variations caused by day-night and seasonal cycles, using methods like SeqSLAM [39] and SMART [40]. The follow up work in this direction is comprised of a number of methods that deal with camera velocity sensitivity [8], [41] or velocity estimation [42], [43]. More recent work includes using temporal information and diffusion process within graphs [44], multi-sequence maps based VPR [45], and trajectory attention-based learning for SLAM [46]. In this paper, we use a simplified sequence-based matching technique mainly to analyse performance dynamics in using
we follow an alternative approach to place representation with each other.

In scene appearance. In particular, measuring the change traverses of the environment, despite significant variations changes are both unique and consistent across multiple places are observed over time. We hypothesize that these information by measuring the change and propose Delta Descriptors that leverage the sequentialtering performed over place-matching scores. In this paper, typically followed by sequential matching or temporal fil-

solves the problem of recognition in a similar way as the method is based on descriptor difference and in essence

of moving objects in order to detect humans. Our proposed actions. [47] used difference subspace analysis to maximize

GDS to tensors for representing and classifying gestures and actions. [19] proposed a novel discriminant analysis based on

a difference vector for analyzing shape differences between objects. [18] proposed Generalized Difference Subspace (GDS) as an extension of a difference vector for analyzing shape differences between objects. [19] proposed a novel discriminant analysis based on GDS demonstrating its utility as discriminative feature extractor for face recognition. Recently, [20] extended the concept of GDS to tensors for representing and classifying gestures and actions. [47] used difference subspace analysis to maximize inter-class discrimination for effective face recognition as an alternative approach to improving representation ability of samples. [48] proposed a human action recognition method based on difference information between spatial subspace of neighboring frames. [22] used sum of depth difference between consecutive frames to discriminate moving/non-

moving objects in order to detect humans. Our proposed method is based on descriptor difference and in essence solves the problem of recognition in a similar way as the GDS-based methods. In particular, Delta descriptors enable dealing with the offset that occurs in the deep-learnt place representations when a robot is operating in new environments under significantly different environmental conditions.

## III. PROPOSED APPROACH

A vast majority of existing place representation methods use single-image based descriptors for place recognition, typically followed by sequential matching or temporal filtering performed over place-matching scores. In this paper, we follow an alternative approach to place representation and propose Delta Descriptors that leverage the sequential information by measuring the change in descriptor as different places are observed over time. We hypothesize that these changes are both unique and consistent across multiple traverses of the environment, despite significant variations in scene appearance. In particular, measuring the change inherently ignores the data bias of the underlying deep-learnt image descriptors, elevating the latter’s utility under diverse environment types and appearance conditions.

In this section, we first highlight key observations from the time-series of existing state-of-the-art image descriptors, then define and formulate the Delta Descriptors, and finally, describe an alternative convolutions-based approach to compute the Delta Descriptors more efficiently.

**a) Key Observations:** In the context of VPR for a mobile robot, images are typically captured as a data stream and converted into high-dimensional descriptors. We consider this stream of image descriptors as a multi-variate time-series. For some of the descriptor dimensions\(^1\) Figure 2 shows pairs of time-series for first 600 images (with 0.5 meters frame separation) from two different traverses of Oxford Robotcar dataset captured under day and night time conditions respectively. For the Raw descriptors, it can be observed that the consecutive values in the time-series tend to vary significantly even though the adjacent frames have high visual overlap. Moreover, the local variations in the descriptor values are not consistent across the traverses albeit the global time-series patterns appear repeatable. As the underlying deep-learnt image descriptors (in most cases) are not trained to be stable against slight perturbations in camera motion or for ignoring the dynamic objects in the scene, such local variations are an expected phenomenon.

**b) Defining Delta Descriptors:** With these observations, we define delta descriptor, \(\Delta_t\), as a signed vector of change measured across a window of length \(l\) over a smoothed multi-variate time-series, \(X_t \in \mathcal{R}^D\), where \(t\) represents the time instant for an observation of a place along a route in the form of an image descriptor. To be more specific, we have

\[
\Delta_t = X_{t+1/2} - X_{t-1/2}
\]  

where \(X_t\) represents the smoothed signal obtained from a rolling average of the time series \(X_t\):

\[
\bar{X}_t = \frac{1}{l} \sum_{t' = t - l/2}^{t + l/2} X_{t'}
\]  

The middle and the right graphs in Figure 2 show the smoothed time-series and the corresponding Delta Descriptor respectively. It can be observed that the proposed descriptors are much more aligned than the baseline ones, getting rid of the offset in their original values.

**c) Simplified Implementation:** The formulation for Delta Descriptors presented above is suitable for understanding and visualizing the time-series patterns. However, Equation 1 and 2 can be simplified to a convolution-based calculation of the proposed descriptors:

\[
\Delta = X \ast W
\]  

where convolutions are preformed along the time axis of the baseline descriptor, independently per dimension using a

\(^1\)The dimension indices were selected using the method described in Section VI-A using NetVLAD descriptors.
1D convolutional filter $\mathbf{W} = (w_1, w_2, \ldots, w_L)$ defined as a vector of length $L = 2l$:

$$w_i = \begin{cases} -1/l, & \text{if } i \leq L/2, \\ 1/l, & \text{otherwise}. \end{cases} \quad (4)$$

For performing visual place recognition, the proposed descriptors are matched using cosine distance. However, for visualization purposes as in Figure 2, individual descriptors are l2-normalized, knowing that the Euclidean distance between pairs of normalized descriptors is proportional to the cosine distance between their un-normalized counterparts.

IV. EXPERIMENTAL SETUP

A. Datasets

We used subsets of two different benchmark datasets to conduct experiments: Oxford Robotcar [17] and Nordland [49]. Repeated route traverses from these datasets exhibit significant variations in scene appearance due to changes in environmental conditions caused by time of day and seasonal cycles.

a) Oxford Robotcar: This dataset is comprised of 10 km traverses of urban regions of Oxford city captured under a variety of environmental conditions. We used the forward-facing camera imagery from the first 1 km of three different traverses, referred to as Summer Day, Winter Day and Autumn Night in this paper. For all three traverses, we used a constant frame separation of 0.5 meters, leading to a database of around 2000 images per traverse.

b) Nordland: This dataset comprises 728 km train journey across vegetative open environment from Nordland captured under four seasons. We used the first 1750 images (out of 35768) from the Summer and Winter traverse after skipping the first 250 frames where the train was stationary.

B. Parameter: Sequence Length

The concept of Delta descriptors is based on measuring the changes in visual information in the form of places observed during a traverse which are then expected to be preserved across subsequent traverses. When using a very short sequence length, such changes can be erratic due to high visual overlap between adjacent frames. This is partly due to the unstable response output from the underlying CNN as it is not trained to produce smooth variation in descriptor for small variations in camera motion, as shown in Figure 2. In order to choose the sequence length parameter for our experiments, we used the relative distribution of cosine distance between descriptors, obtained by matching a dataset with itself.

Figure 3 shows these distributions for Winter Day traverse from the Oxford Robotcar dataset and Summer traverse from the Nordland dataset, where cosine distance is plotted against frame separation as the median value computed across the whole traverse. We found that using a fixed cosine distance threshold of 0.7 (black horizontal line), a minimum bound on the sequence length parameter can be directly estimated from these distributions (shown with red circles), making sure that the place observation has changed sufficiently enough to robustly measure the changes in descriptor values. Using this method, the sequential span was found to be 38, 36 and 57 for the Oxford datasets: Winter Day, Summer Day and Autumn Night. For the Nordland Summer and Winter traverses, these values were estimated to be 14 and 8. Hence, in our experiments, we compute Delta Descriptors using a fixed sequence length of 64 and 16 frames for all the traverses of the Oxford Robotcar and the Nordland dataset respectively.

C. Evaluation

We used Precision-Recall (PR) curves to measure VPR performance. For a given localization radius, precision is defined as the ratio of correctly matched queries to total retrieved queries and recall is defined as the ratio of correctly matched queries to total number of queries with a defined ground truth match, where a match for a query is retrieved only when its cosine distance is less than a threshold which is varied to generate the PR curves. We present PR curves for two different values of localization radii: 10 and 40 meters for the Oxford Robotcar dataset and 2 and 10 frames for the Nordland dataset.

D. Comparative Study

We used the state-of-the-art single image-based descriptor NetVLAD [11] as a baseline in the results, represented as Raw Descriptors. Delta Descriptors were calculated using Equation 3 with NetVLAD\footnote{The proposed method is not tied to this underlying descriptor; the experiments in Section VI-A were conducted using a ResNet-based GMP descriptor.}. As the proposed Delta descriptors use sequential information, we also compare them against a naive sequential representation of NetVLAD, achieved by smoothing the baseline descriptors using Equation 2 represented as Smoothed Descriptors. Furthermore, we also consider the orthogonal approach to utilizing sequences for VPR that is based on sequential aggregation of match scores, typically obtained by comparing single image descriptors. For this, we use a simplified version of sequence matching which is similar to [39] but only aggregate match scores along a straight line without any velocity searching [50]. We refer to this as SeqMatch in the results and use it on top of the Raw, Smoothed and Delta descriptors.

V. RESULTS

In this section, we first present the benchmark comparisons on three pairs of route traverses using two different datasets.
Then, we demonstrate the performance effects of PCA variations occur in natural open environment (Figure 4b). Such cannot achieve performance similar to that attained using sequence matching on top of raw descriptors (solid blue) to mitigate the issue in a completely unsupervised manner. Supervised fine-tuning or re-training, Delta Descriptors tend of the NetVLAD descriptors. Thus, in contrast to requiring low performance might be due to lack of generalization ability descriptors is observed to be quite low when such appearance environment (Figure 4a), the absolute performance of raw performance levels are observed to be saturated for the day-

A. Sequential Representations and/or Sequential Matching

Figure 4 shows the results for two pairs of traverses from the Oxford Robotcar dataset and one pair from the Nordland datasets. It can be observed that Delta Descriptors perform better than the Raw and Smoothed descriptors in all the comparative studies. Furthermore, state-of-the-art performance is achieved when sequence matching is used on top of Delta Descriptors.

B. Dimension Reduction via PCA

Image descriptors obtained through CNNs are typically high-dimensional. For global retrieval tasks, computational complexity is often directly related to the descriptor dimension size. Therefore, dimension reduction techniques like PCA are commonly employed [11], [53], [54], [55]. However, this can lead to significant performance degradation due to extreme variations in scene appearance as the variance distribution in the original descriptor space may not be repeatable. In Figure 5a, we show the effect of PCA-based dimension reduction on the performance of Raw NetVLAD and Delta descriptors using the Oxford Robotcar day-night traverses. It can be observed that the proposed Delta Descriptors are robust to dimension reduction techniques like PCA: even retaining only 50 principal components does not degrade the recall performance much. On the other hand, the baseline NetVLAD descriptors suffer significant performance drop with PCA even when all 4096 components are retained, highlighting its sensitivity to data centering.
C. Order of Place Observations

The concept of Delta descriptors is based on the sequential order of changes in visual information. In the context of VPR, sequentially observed places typically have some visual overlap which affects the overall dynamics of performance when considering either sequential representation or sequential matching. We consider another scenario where both the reference and the query data are shuffled such that the order of images is preserved across traverses but there is no visual overlap between adjacent frames. For this, we used the Nordland dataset and sampled every $100^{th}$ image out of 35768 images and then performed the shuffling. In Figure 5, we can observe that even with a sequential span of 2 frames (lacking visual overlap) and localization radius of 1 frame, additional information in the form of sequences can be better utilized with sequence-based descriptors than sequence-based match-score aggregation of single image descriptors, while their combination achieves even higher performance. Furthermore, this experiment also indicates that the concept of Delta Descriptors is applicable in general to describing and matching ordered observations, irrespective of whether or not the adjacent elements are related to each other.

D. Camera Motion Variations & Multi-Delta Descriptors

In our previous experiments, we used a constant frame spacing between the reference and the query traverses (0.5 meters for Oxford Robotcar). In practice, camera velocity may change both within and across the repeated traverses of the environment. In order to observe the effect of such variations on the VPR performance, we conducted another experiment using the first 8600 image\(^4\) (\(\sim\)1 km) from the Winter Day and Autumn Night traverses without any data pre-processing, that is, motion-based keyframe selection.

For this study, we used a sequential span of 64 frames both for computing Delta Descriptors and sequence matching. Furthermore, in order to deal with variable motion both across and within the traverses, we present a preliminary investigation into sequential-span searching using a Multi-Delta Descriptor approach. To achieve this, for both the reference and the query data, multiple Delta descriptors are computed using a range of sequential spans: \{30, 40, 50, 60\} frames in this case. The match value for any given image pair is calculated as the minimum of the cosine distance over all the possible combinations of sequential spans (16 here) used for computing multiple Delta Descriptors.

In Figure 6, we can observe that even without any data preprocessing (keyframe selection), state-of-the-art performance is achieved by the Delta Descriptors + SeqMatch combination. As we do not use any local velocity-search technique (originally proposed in [39]) in our sequential matching method, performance contribution of SeqMatch is less due to velocity variations as compared against the constant-velocity experiments in Figure 4. However, the effect of local variations in camera velocity is less detrimental for Delta descriptors, leading to superior absolute performance even without any sequential matching. Finally, it can be observed that the Multi-Delta Descriptor approach further improves the state-of-the-art performance. This also emphasizes the highly discriminative nature of difference-based description that enables accurate match searching within a given range of sequential spans, potentially leading to its applications beyond exact repeats of route traversals.

VI. DISCUSSION

A. Visualizing Activations

In order to visualize how Delta descriptors utilize the difference-based sequential information, we used a Global Max Pooling (GMP) based image descriptor through which image region activations can be directly interpreted which is not trivial for VLAD pooling of NetVLAD. For this purpose, a GMP descriptor was extracted from the final conv layer of ResNet-50 [56] and Delta Descriptors were calculated using GMP descriptors for Oxford’s Winter Day and Summer Day traverses. Figure 7 shows performance comparison between the Raw and Delta descriptors using GMP as the underlying descriptor. It can be observed that the performance of Delta Descriptors is directly related to the underlying descriptors as for the same pairs of traverses, recall performance is significantly better when using NetVLAD (see Figure 4).

For visualization purpose, using the ground truth for place matches, dimensions of the GMP descriptor were ranked in order to observe only those which contributed the most to the performance. This was achieved by taking an element-wise product of a known matching pair of descriptors and sorting

---

\(^4\)only every $5^{th}$ frame was considered, leading to the dataset size of 1720 images each. Note that this does not affect the camera velocity and was only done to reduce the processing time.
them in the order of the product value (a higher value indicates that both the descriptors had a similar high activation). This process was repeated for all the pairs of descriptors (~2000) and the dimensions that repeatedly ranked higher (higher product value) were selected for visualization.

In Figure 7(a) and (c), GMP descriptor dimension index 1501 is used. The graph in Figure 7 shows the variation of descriptor values along the route for Raw, Smoothed, and Delta descriptor (from top to bottom). For image indices in the range of 400 – 800, both the raw and the smoothed values do not align well across the traverses but are relatively closer in the Delta Descriptor space. Figure 7 displays the 550th image index from the Winter (left) and the Summer day traverse (right) where mask color indicates the activation values that increase from blue to green and then to red. It can be observed that the activations for the image from the Summer traverse are lower than that for the Winter traverse due to different lighting conditions around the visual landmark, leading to an increased distance in the original descriptor space. However, as the Delta descriptor only considers changes within a traverse, its descriptor values still remain consistent throughout, even though the absolute activation values are lower in one of the traverses.

VII. CONCLUSION AND FUTURE WORK

Visual place recognition under large appearance changes is a difficult task. Existing deep-learned global image description methods do enable effective global image retrieval. However, when operating in new environments where appearance conditions vary drastically due to day-night and seasonal cycles, these methods tend to suffer from an offset in their image description. Our proposed Delta Descriptors are defined in a difference space, which is effective at eliminating description offsets seen in the original space in a completely unsupervised manner. This leads to a significant performance gain, especially for the challenging scenario of day-night

VPR. When considering a given sequential span, we have demonstrated that Delta Descriptors achieve state-of-the-art results when combined with sequential matching. This performance is a strong indicator of the robust representation ability given by Delta Descriptors. Finally, we have presented results for a range of experiments that show the robustness of our method when handling PCA-based dimensional reduction and variations in camera motion both along and across the repeated route traversals.

Our current work can be extended in several ways including estimating the descriptor transformation (offsets) on the fly, learning what visual landmarks are more suited to track changes, and measuring changes independently but simultaneous for different descriptor dimensions. In particular, it would be interesting to see how a framework that learns underlying descriptors would change its behaviour if optimized for place recognition performance using the subsequent Delta Descriptors. The concept of using a difference space is not well-explored in the place recognition literature but is a promising avenue for future research applied to other similar problems where inferring or learning the changes might be more relevant than the representation itself. We believe that our research contributes to the continued understanding of deep-learned image description techniques and opens up new opportunities for developing and learning robust representations of places that leverage temporal information.

REFERENCES

[1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.
[2] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in Computer Vision (ICCV), 2011 IEEE international conference on. IEEE, 2011, pp. 2564–2571.
[3] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 886–893.
[4] A. Oliva, “Gist of the scene,” Neurobiology of attention, vol. 696, no. 64, pp. 251–258, 2005.
[5] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in Proceedings of International Conference on Computer Vision (ICCV). IEEE, 2003, p. 1470.
[6] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation,” in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 3304–3311.
[7] M. Milford, “Visual route recognition with a handful of bits,” Proc. 2012 Robotics: Science and Systems VIII, pp. 297–304, 2012.
[8] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, “Robust visual robot localization across seasons using network flows,” in Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.
[9] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, “Lift: Learned invariant feature transform,” in European Conference on Computer Vision. Springer, 2016, pp. 467–483.
[10] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, “Large-scale image retrieval with attentive deep local features,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3456–3465.
[11] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad: Can architecture for weakly supervised place recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297–5307.

5 All the descriptors are l2 normalized independently to aid visualization in Euclidean space.
