machines

MDPI

Article

Visual Place Recognition in Changing Environments with Sequence Representations on the Distance-Space Domain

Ioannis Tsampikos Papapetros [1,*], Ioannis Kansizoglou [1], Loukas Bampis [2] and Antonios Gasteratos [1]

[1] Department of Production and Management Engineering, Democritus University of Thrace, Vas. Sophias 12, GR-671 32 Xanthi, Greece; ikansizo@pme.duth.gr (I.K.); agaster@pme.duth.gr (A.G.)
[2] Department of Electrical and Computer Engineering, Democritus University of Thrace, Building B, Kimmeria Campus, GR-671 32 Xanthi, Greece; lbampis@ee.duth.gr
* Correspondence: ipapapet@pme.duth.gr; Tel.: +30-2541-079359

Abstract: Navigating in a perpetually changing world can provide the basis for numerous challenging autonomous robotic applications. With a view to long-term autonomy, visual place recognition (vPR) systems should be able to robustly operate under extreme appearance changes in their environment. Typically, the utilized data representations are heavily influenced by those changes, negatively affecting the vPR performance. In this article, we propose a sequence-based technique that decouples such changes from the similarity estimation procedure. This is achieved by remapping the sequential representation data into the distance-space domain, i.e., a domain in which we solely consider the distances between image instances, and subsequently normalize them. In such a way, perturbations related to different environmental conditions and embedded into the original representation vectors are avoided, therefore the scene recognition efficacy is enhanced. We evaluate our framework under multiple different instances, with results indicating a significant performance improvement over other approaches.

Keywords: visual place recognition; changing environments; sequence matching; localization; navigation

Citation: Papapetros, I.T. ; Kansizoglou, I.; Bampis, L.; Gasteratos, A. Visual Place Recognition in Changing Environments with Sequence Representations on the Distance-Space Domain. *Machines* **2023**, *11*, 558. https://doi.org/10.3390/machines11050558

Academic Editors: Jian Wu, Xiangkun He and Guangfei Xu

Received: 27 March 2023
Revised: 5 May 2023
Accepted: 11 May 2023
Published: 16 May 2023

1. Introduction

The utilization of autonomous mobile robots promises to revolutionize numerous industrial and domestic fields. Navigating safely through the environment is of utmost importance for the majority of those applications. Yet, at the center of every navigation module lies a localization framework, the purpose of which is to determine the platforms' position relative to their environment. In the vast variety of related approaches, visual localization [1] holds a vital role and can offer an edge over the alternatives [2–4], depending on the nature of the underlying problem.

The task of visual place recognition (vPR) describes the process of searching a visual scene into a database of mapped places in order to locate a match, should one exist [5]. Multiple methods have been proposed to address this task, offering adequate results [6]. However, in most cases, their employment in a continuously running scenario significantly limits their performance both in terms of scalability [7] as well as the change in the environment due to the difference in the visual conditions [8] or the existence of dynamic objects [9]. In spite of applications where the observable environment can be assumed to have a constant state, long-term practices should be able to recognize places under severe environmental changes [10–12].

With this article, we propose an unsupervised long-term visual place recognition technique which utilizes sequential observations to reinforce the individual scene representations. We argue that those individual representations are not a direct depiction of the content of the environment, but are heavily affected by the observed visual conditions. The aim of this work is to alleviate the aforementioned condition effect and thus improve the performance of vPR. This is achieved by remapping the image encodings from the

descriptor-space to the distance-space and subsequently normalizing the produced signals. In such a way, any existing constant noise between the underlying representations is removed prior to the matching process. Thus, the scene's structure is captured in a similar manner between the database and query observations, alleviating the effects of environmental changes. More specifically, the main contributions of this work can be summarized as follows:

1. Introducing a process for data manipulation in the distance-space domain, with a view to solve the vPR problem;
2. Offering a concept for a descriptor-agnostic procedure for long-term vPR practices, utilizing multiple sequential observations and generalizing place-specific representations into different and unseen environments;
3. Providing insights about the utilization of image sequences for scene representations;
4. Using multiple datasets to evaluate the behavior of the proposed method under different situations, such as velocity and viewpoint fluctuations, varied environments, as well as observed conditions, etc.

In the following section we briefly analyze the related literature, wherein we situate this work. In Section 3, an intuitive explanation of this framework is presented, followed by an extensive description of its implementation details. Subsequently, Section 4 features the method's evaluation procedure, while Section 5 presents the conclusions, a discussion of the key findings and a prompt for possible extensions.

2. Related Work

A vPR framework should be able to efficiently map the observed environment, so that later impressions of a place can be easily identified and retrieved, despite any condition-related changes in its appearance. During navigation, a robot constantly perceives its environment via repetitive discrete sensor readings. Those discrete readings—in this article considered as images—can be used as a source of information relative to the place they were observed. Representing a visual scene with encodings of that information is a fairly popular technique in the vPR challenge [6,7].

Employing local features, i.e., individual patches within the image [13], one can detach the extracted information from the underlying viewpoint-related spatial structure of the scene, enabling the matching of different poses of the same content observations [14–17]. On the other hand, utilizing the spatial relationship of the extracted information on the image plane may limit the viewpoint tolerance but can significantly improve the matching quality of places with major condition-related differences [8,18–20].

While the single-view scene encodings can provide a natively simple vPR method with satisfactory results, augmenting the input data with information from temporally close observations can significantly improve the final outcome. A way of utilizing the sequential information of an observation is to reinforce or weaken the belief that a matched place is a true positive detection. This can be achieved by observing whether or not temporally close observations are matched to the same place or are scattered in the database [11,21,22]. This technique (sequence-based matching) is highly effective, especially in cases prone to perceptual aliasing, and it can be used as an extra layer to filter out uncertain detections.

Furthermore, appending scene information from sequential and discrete observations into a single representation vector, or exploiting the spatial relationship of the visual clues beyond the two-dimensional plane, so as to form the sequence-based representations of a place, can enhance the amount of the extracted information [23–26]. However, merging sequential data without some kind of supervision can introduce noise, degrading the place recognition performance. Thus, several methods have adopted a sequence generation module to filter out noisy visual clues [21,27].

In a similar manner, our method utilizes multiple successive frames to reinforce the perception of the observable scene. Although the lack of a sequence generation module unavoidably introduces some kind of noise, the proposed framework is able to achieve state-of-the-art performance.

3. Method

3.1. Intuition

The purpose of place recognition is to map the environment and speculate whether or not a scene has been seen in the past. To achieve that, each scene has to be somehow abstracted. In this subsection, we will attempt to intuitively analyze the overall pipeline of describing a visual scene in place recognition.

Ideally, we would like to measure the similarity of the scenes based on their content. Unfortunately, the available scene perception methods fail to identify the whole of the underlying information. Furthermore, the process of capturing this information, the condition of the environment, as well as the utilized encoding method can significantly affect the produced representation. Building on that intuition, our method aims to remove the condition effect from the final similarity result of the matching between two scenes. This is achieved by remapping the representation vectors from the description-space into the distance-space, while appropriately representing them, so as for matching assessment to be feasible (Figure 1). In this context, the distance-space is the domain where observations are represented as their distance from a given target. Thus, instead of directly measuring the similarity between scenes, we utilize the relative distance of sequential readings from a common reference. Typically, the calculated relative distance among a specific pair of representation vectors could be generated by a multitude of different vectors, thus decreasing its distinctiveness. In contrast, using sequences of those relative distance values generates highly improbable combinations to be observed randomly. This way, a mutual frame of reference is provided allowing a fair matching. With the reference image belonging into the first 'query' scene, the produced distance vector of the second one can be intuitively perceived as a similar vector displaced by an approximately constant coefficient. This coefficient originated from the difference in environmental conditions between the query frame and the second scene, and it is not present in the first scene's vector. Therefore, a normalization scheme is employed to remove this component, discarding the condition effect prior to the matching procedure.

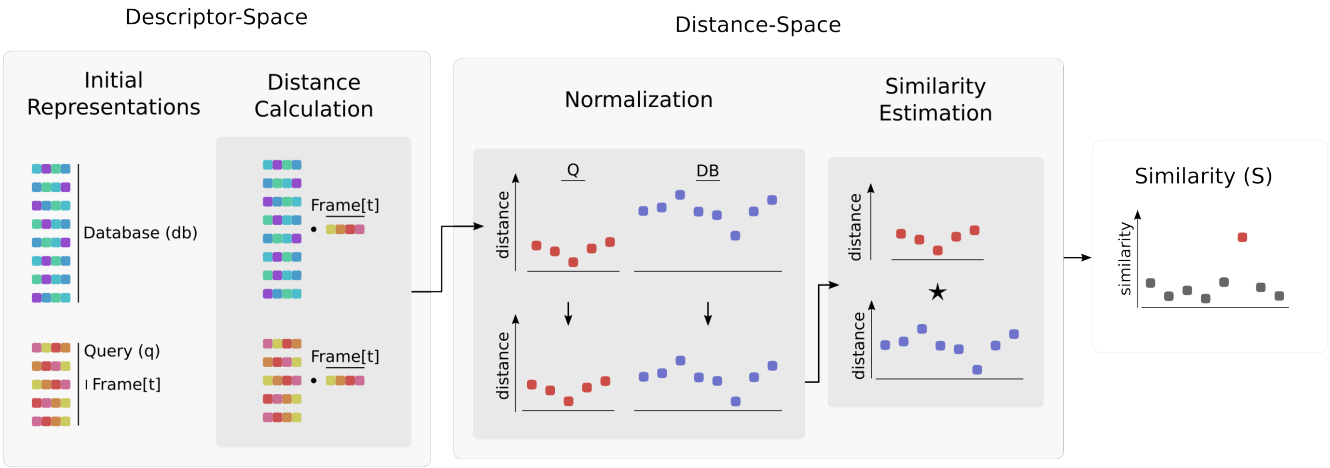


Figure 1. An overview of the proposed method. First, the distances between the representations of the frame of interest and both the query (q) and database (db) sequences, are calculated, forming the new distance-space representations Q and DB. Then, the normalization process takes place, regularizing the produced distance signals to lie within the same range. Finally, the similarity among the frame at timestep t and the database is estimated by sliding the newly formed Q signal over the normalized DB one and computing their cross-correlation (star symbol).

3.2. Implementation

As already mentioned, in order to effectively compare visual scenes, some sort of representation vectors should be generated from the captured images. Considering the

marginal effect of the employed representation method in our work, without loss of generality, we assume that the image representation vectors have been already generated.

Let us define a given database (db) as a stream of all seen scenes represented by multidimensional vectors, with l_{db} indicating the length of db. In addition, we consider a query trajectory as a similar stream, from which we form the query sequences (q), by selecting a vector of the image of interest concatenated with its adjacent ones. Given the above, we transform both the database and the query sequences into a one-dimensional time-series. In specific, considering the image of interest at time t, we compute its distance from each image in db as well as each corresponding one in q. Using the following formulas, the new query $Q = [\ldots, Q_i, \ldots], \forall i \in \{i \in \mathbb{N} : t - w \leq i \leq t + w\}$ and database $DB = [\ldots, DB_j, \ldots], \forall j \in \{j \in \mathbb{N} : 0 \leq j \leq l_{db}\}$ signals result from:

$$Q_i = 1 - \frac{\hat{q}_i \cdot \hat{q}_t}{||q_i|| \cdot ||q_t||}, \forall i \in \{i \in \mathbb{N} : t - w \leq i \leq t + w\}, \quad (1)$$

$$DB_j = 1 - \frac{\hat{db}_j \cdot \hat{q}_t}{||db_j|| \cdot ||q_t||}, \forall j \in \{j \in \mathbb{N} : 0 \leq j \leq l_{db}\}, \quad (2)$$

where w represents half the size of the sliding window containing the adjacent images in the query stream. The upper plot in Figure 2 shows a snapshot of the produced one-dimensional signals on a sample subset of the Nordland [12] dataset. It can be observed that the database signal tends to match the query one, displaced by a constant component.

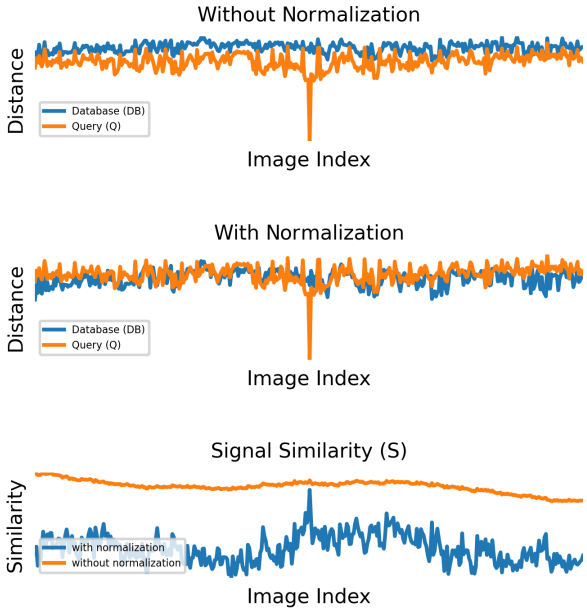


Figure 2. A formed query (Q) representation signal along its database (DB) correspondence on the Nordland dataset [12], before (upper row) and after (middle row) the normalization procedure. Using Q as a sliding window over DB, the cross-correlation results (lower row) of those signal pairs demonstrate that the normalized ones, produce a crisp similarity peak, vigorously indicating a query match in the database.

By utilizing the process of template matching, i.e., calculating the similarity of the template Q over DB, one can deduce meaningful information regarding the origin of the query signal in the database. Although, due to the non-standardized amplitude of those signals, the resulting similarity can be severely affected (Figure 2). To that end, a z-score normalization [28] is applied over both signals, to express the underlying data relative to their mean and standard deviation values, thus eliminating the observed displacement on the amplitude (vertical) axis, as shown in the middle plot of Figure 2. The z-score function is defined as:

$$zs(x) = \frac{(x - \mu)}{\sigma}, \tag{3}$$

where x represents the data, whilst μ and σ are the mean value and the standard deviation of that data, respectively.

Finally, the process of identifying the similarity (S) between the query Q and the database DB signals is performed by calculating the correlation of the normalized Q with the DB representations. In particular:

$$S = zs(DB) \star zs(Q), \tag{4}$$

with $\star$ notating the cross-correlation operation between the two signals. The output of (4), as well as the matching result of the equivalent non-normalized DB and Q signals are shown in the lower plot of Figure 2.

4. Evaluation

A robust place recognition framework should be able to recognize scenes under severe environmental changes. In this section, we assess the behavior of the proposed method in several critical aspects of such a system.

4.1. Datasets

1. *The Nordland dataset:* to evaluate the framework's resilience in recognizing visual scenes under different seasonal conditions, the Nordland dataset [12] was employed. Its content features front-facing images of a 729 km long railway journey, captured in four seasons. Moreover, due to the rail-moving camera, this dataset minimizes the viewpoint variations between each trajectory. For this evaluation, we selected the test partitions [8] of the summer and winter trajectories, each comprising of 3450 synchronized frames.
2. *The Oxford dataset:* in contrast to the one-to-one image correspondence of the Nordland tracks, the Oxford RobotCar dataset [10,29] contains routes from the central Oxford with inconsistent velocities and viewpoints. The data were captured by a car-mounted sensor during different conditions. From the available data, we chose the central view of the front-facing camera of the day and night trajectories (parts 01 of 2015-02-17-14-42-12 and 2014-12-16-18-44-24, respectively). Moreover, due to faulty GPS signal, we discarded the first 2000 frames, resulting in two routes, each one consisting of 4000 images. Within the trajectory, the mean traveled distance between two successive frames is 0.26 m.
3. *The COLD dataset:* unlike the above-mentioned datasets, the COLD collection (sequence 1 of the Freiburg set) [30] comprises of indoor routes, depicting an office-like environment at different time periods. Similar to the Oxford dataset, its trajectories are asynchronous and they exhibit several viewpoint mutations. Within this work, we selected the sunny and night instances, containing 1598 and 1911 images, respectively, with a mean step distance equal to 0.04 m.

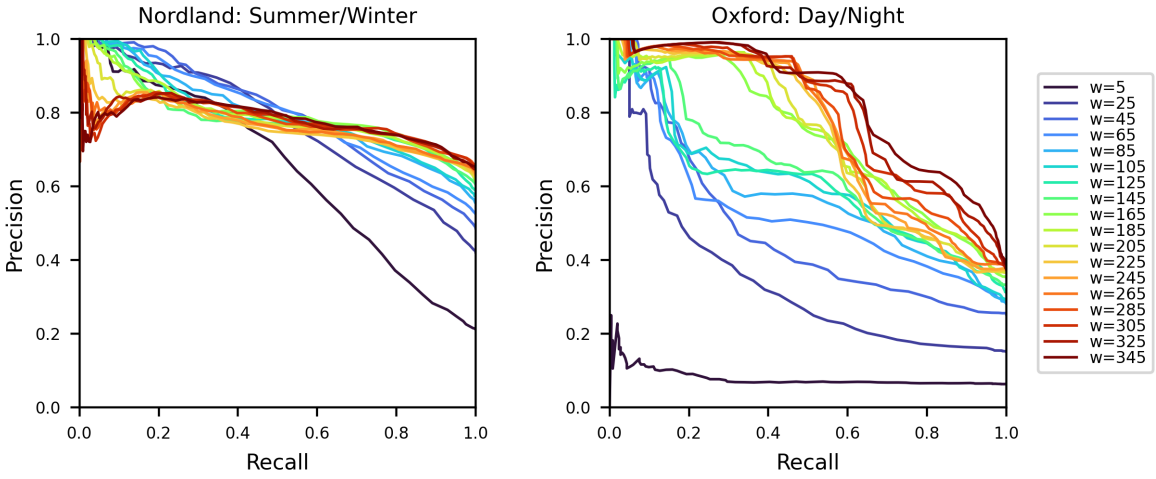
From the above three cases, the COLD dataset was not considered during the evaluation of our proposal's behavior over different parameterizations since its respective results were less informative. However, it is considered in order to provide a distinct scenario for our assessment when comparing against the state-of-the-art.

4.2. Metrics

To examine the performance of our approach, we need to quantify and compare the produced results. Among the plethora of available metrics, in this section we utilize precision-recall and their area under curve (AUC) metrics [31].

4.3. The Parameter w

As described in Section 3.2, the length of the query observation depends on the parameter w, which denotes half the window size (in terms of the number of images) around frame t. Figure 3 contains the precision and recall curves we obtained, by varying a similarity threshold over score S for different values of w on the Nordland and Oxford datasets. The employed image representations for this experiment were generated using the NetVLAD method [18]. Whilst changes in w do not seem to significantly affect the framework's performance on the Nordland routes (Figure 3a), higher values tend to produce better results on the Oxford ones (Figure 3b). This can be explained as a consequence of the non-synchronous frames acquisition of the latter dataset among its trajectories. More specifically, comparisons among highly dissimilar places, i.e., distant from time t, introduce noise during the cross-correlation operation in (4), degrading the final outcome. Yet, larger sequences can accumulate additional information about the observed place and thus reinforce the underlying representations, leading to enhanced performance. Furthermore, as mentioned in Section 3.1, representing an observation using a single value offers trivial distinctiveness. On the contrary, aggregating multiple values from a sequence of observations produces a highly distinctive combination that leads to greater matching performance. Indeed, the results confirm this prediction, as in both datasets higher w values tend to produce better results.



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

Figure 3. The resulting precision and recall curves of our method on the Nordland (**a**) and the Oxford (**b**) datasets, for various values of parameter w. The utilized ground truth radius is $R = 10$ and $R = 40$ for Nordland and Oxford datasets accordingly.

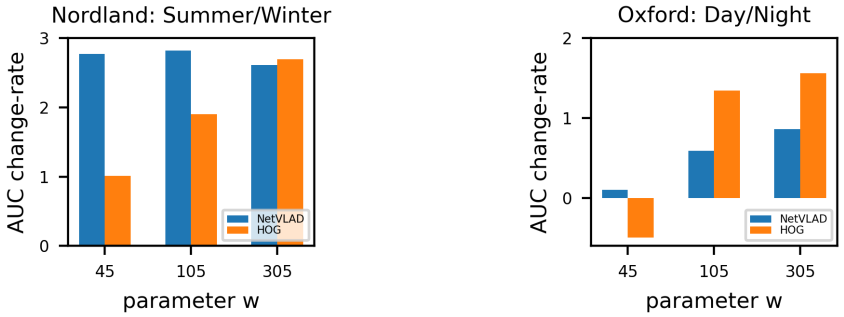
4.4. Different Descriptors

To evaluate the descriptor-agnostic comportment of our framework, we tested a couple of different image encoders. More specifically, we employed the NetVLAD encoder [18], which constitutes a popular well-established network choice for vPR tasks [32,33] and the histogram of oriented gradient (HOG) [34] technique, i.e., a classical unsupervised method utilized by many contemporary systems in the field [17,35,36]. Figure 4 shows the performance of the proposed method, in terms of the change-rate of the AUC score over the raw representations, while feeding different descriptor types as an input. The AUC change-rate is calculated using:

$$cr = \frac{s - r}{r}, \tag{5}$$

with s indicating the AUC score of our method. The values r were obtained by varying a threshold over the similarity metrics among NetVLAD and HOG descriptors, respectively, and computing the corresponding AUC scores. In this manner, the AUC change-rate

indicates the performance improvement provided by our technique regardless of the underlined description approach, with higher values of the w parameter consistently performing much better than the baseline.



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

Figure 4. Performance evaluation over different representation methods, with the applied metric being the AUC score change-rate of our method over the raw descriptors' one in (**a**) the Nordland and (**b**) the Oxford datasets. The used ground truth radius is $R = 10$ and $R = 40$ respectively.

4.5. Sequence Shuffling

When revisiting a place, images of the environment do not always appear in the same order as its initial observation. In this subsection, we introduce a case study wherein we randomly shuffle the observed images during the query time to emulate such an event. This operation adds noise to the similarity estimation of (4), thus degrading the overall performance. For this case study, we chose to run the experiments on the Nordland dataset due to its one-to-one frame correspondence, as well as the consistent viewpoint of the captured images. Moreover, we fix the parameter w at an average performing query window ($w = 105$). In Figure 5, we plot the AUC score of such experiments, using different percentages of shuffled images from the query sequence. As presented, despite the high percentage of randomly shuffled images within the query observation, the proposed method manages to maintain great results, while outperforming the raw descriptors' score.

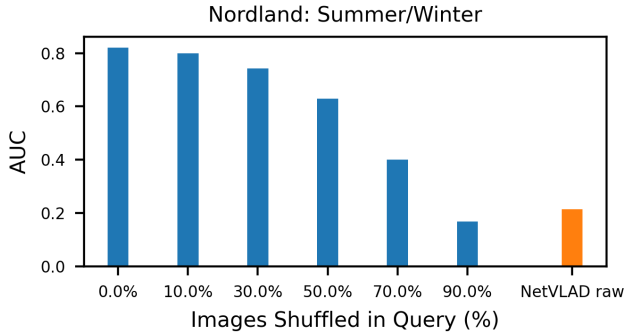


Figure 5. The AUC score of the proposed method, while randomly shuffling a different percentage of the images within the query sequence, alongside the single-image NetVLAD descriptors' result. The applied ground truth radius is $R = 10$. Despite the large percentage of shuffled images, our method manages to maintain great results, outperforming the raw descriptors' score.

4.6. Execution Time

The significance of the place recognition module in the robot navigation pipeline places it under real-time execution constraints. Table 1 shows the mean execution time per image of every part of our system while processing the Nordland dataset, with NetVLAD-generated encodings. As outlined, our method manages to improve the raw NetVLAD descriptors' performance with minimal overhead. The experiments were performed using a Python-based on an 8-core 7-7700K CPU @4.20 GHz and 32 GB of memory.

Table 1. Execution time per image of our method. The data were captured using NetVLAD-generated image representations from the Nordland dataset.

	$\mu \pm \sigma$ **(ms)**
Distance Calculation	17.76 ± 0.55
Normalization	0.28 ± 0.04
Similarity Estimation	0.26 ± 0.08
Total	18.31 ± 0.68

4.7. Comparison with State-of-the-Art

Whilst the proposed framework demonstrates promising results, it is important to compare them with the performance of similar state-of-the-art techniques, so as to offer an integral understanding regarding its placement within related works in the field. For that purpose, the sequence-based representation method of Delta Descriptors [25] was employed, being the most relevant state-of-the-art approach; alongside the computed raw single-image NetVLAD descriptors, to form the comparison study of this subsection. Figure 6 outlines the benchmark results of the tested methods against the selected datasets, in terms of the precision and recall curves. Similarly, Table 2 exhibits the same results in a single benchmarking value, in terms of AUC score. Note that for the above results, the required place recognition ground-truth was obtained by accepting true-positive matches with the database instance lays within a certain radius R from the query one, in terms of frames for the Nordland and meters for the Oxford and COLD datasets. As shown, the proposed method tends to outperform both the state-of-the-art, as well as the baseline. Moreover, it is worth noting the slightly reduced efficacy relative to the general trend, when using a lower radius for accepting true-positive matches as per the ground-truth. Such a reduction is mainly owed to the multi-frame observations, which are embedded in the used representations, offering a reliable yet imprecise estimate of the correct place in the database. Lastly, the execution time of both methods can be considered minimal, with Delta Descriptors performing slightly slower at 20.47 ms, as reported by deploying the publicly available source codeon the same hardware.

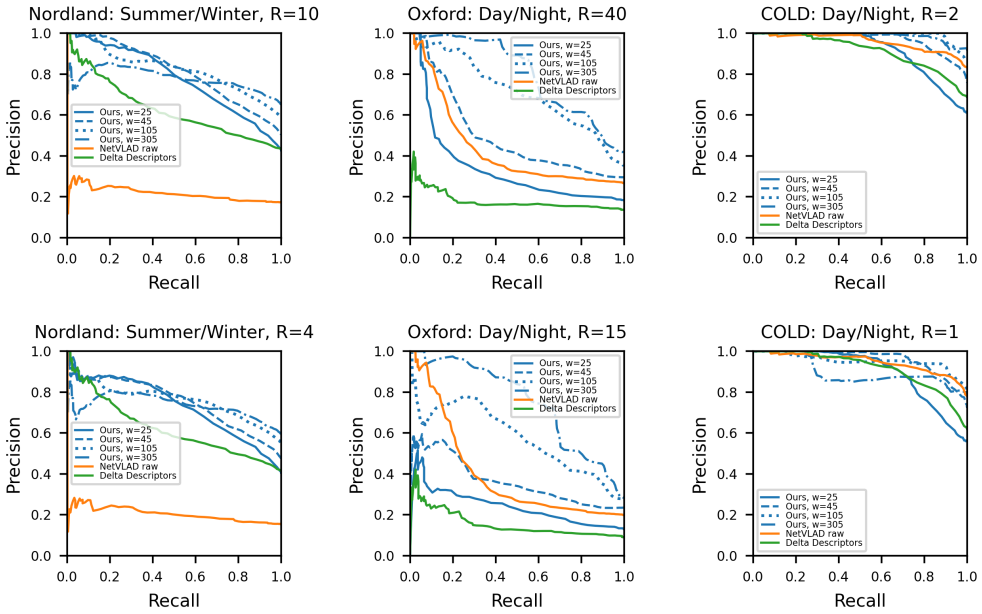


Figure 6. Precision and recall curves of the proposed method, along Delta Descriptors and the raw NetVLAD vectors, in different datasets. Parameter R, indicates the radius of correct matches around the ground-truth, in terms of frames for the Nordland dataset and meters for the Oxford and COLD ones. As shown, our method tends to outperform the compared methods in the tested datasets.

Table 2. The AUC score of the proposed method, along Delta Descriptors and the raw NetVLAD vectors, in different datasets. Parameter R indicates the radius of correct matches around the ground-truth, in terms of frames for the Nordland dataset and meters for the Oxford and COLD ones.

	Nordland $R = 10$	**Oxford** $R = 40$	**COLD** $R = 2$	**Nordland** $R = 4$	**Oxford** $R = 15$	**COLD** $R = 1$
Delta Descriptors	0.626	0.166	0.913	0.621	0.147	0.914
NetVLAD raw	0.215	0.438	0.957	0.203	0.401	0.946
Ours $w = 25$	0.767	0.351	0.914	0.725	0.267	0.898
Ours $w = 45$	0.813	0.499	0.969	0.752	0.367	**0.961**
Ours $w = 105$	**0.823**	0.714	**0.990**	**0.760**	0.580	0.951
Ours $w = 305$	0.779	**0.799**	0.988	0.736	**0.751**	0.896

4.8. Query Window as a Time Trace

So far, we have analyzed the performance of the proposed method while placing the query image in the center of the q sequence. In such a layout, the representation benefits by incorporating images from the trajectory without over-extending to frames far from t, thus limiting the drawback of including noise by matching highly dissimilar places, as described in Section 4.3. Yet, in scenarios where the most recent image needs to be utilized for vPR, the described layout is not feasible. For those cases, we modify the query sequence as $Q_{tt} = [\ldots, Q_i, \ldots], \forall i \in \{i \in \mathbb{N} : t - 2w \leq i \leq t\}$ using:

$$Q_i = 1 - \frac{\hat{q}_i \cdot \hat{q}_t}{||q_i|| \cdot ||q_t||}, \forall i \in \{i \in \mathbb{N} : t - w * 2 \leq i \leq t\}. \quad (6)$$

We conducted the same experiments described in Section 4.7 employing the modified query sequence Q_{tt}. Table 3 shows the results of the time trace window experiment. As observed, the AUC score slightly deteriorates in comparison to the results from Table 2, as well as the top-performing window sizes tend to shift to smaller w values. The presented outcome is expected, as a consequence of the abovementioned limitation regarding the added noise in the query sequence.

Table 3. The AUC score of the proposed method, using the time trace window approach, in different datasets. Parameter R indicates the radius of correct matches around the ground-truth, in terms of frames for the Nordland dataset and meters for the Oxford and COLD ones.

	Nordland $R = 10$	**Oxford** $R = 40$	**COLD** $R = 2$	**Nordland** $R = 4$	**Oxford** $R = 15$	**COLD** $R = 1$
$w = 25$	0.664	0.415	**0.914**	0.602	0.381	**0.753**
$w = 45$	0.642	**0.618**	0.913	0.590	**0.483**	0.686
$w = 105$	0.735	0.497	0.693	0.687	0.372	0.352
$w = 305$	**0.848**	0.577	0.455	**0.825**	0.225	0.384

5. Conclusions

The autonomous operation of mobile robots in highly mutable environments constitutes a great challenge. Severe appearance changes can significantly limit their navigation ability, thus prohibiting their use in many real-world applications. The article at hand proposed a sequence-based unsupervised vPR technique, capable of operating under severe environmental changes, with the aim to support the navigation procedure of autonomous mobile robots in real-life scenarios. In contrast to the single-frame solutions, utilizing multiple observations to form a single representation significantly enhanced the quality of the produced results. Furthermore, by utilizing the distance-space representation and normalization, our method managed to robustly match scenes across different environmental conditions, without explicit training or parameterization.

Although our method manages to produce superior results along a multitude of environments, observed under challenging conditions, a trade-off between quantity and

quality is realized. As described in Section 4.3, the w parameter regulates the number of observations incorporated in the query representation. While higher values of w augment the useful information leading to enhanced matching performance between the query scene and the database, a noise generated from the matching of inconsistent frames amid the associated sequences degrades the final outcome. Furthermore, due to the decreased distinctiveness of the generated distance vectors (Section 3.1), the utilized sequence sizes tend to be quite large. Depending on the specifics of the application, this peculiarity poses a consideration.

As part of future work, we intend to extend our framework with the utilization of an automated sequence generation module. Such a module should suppress the above-mentioned noise, as well as improve the achieved performance for identifying the exact revisited database instance, as explained in Section 4.7.

Author Contributions: Conceptualization, I.T.P. and A.G.; methodology, I.T.P.; software, I.T.P.; validation, I.T.P., I.K. and L.B.; formal analysis, I.T.P. and I.K.; data curation, I.T.P.; writing—original draft preparation, I.T.P.; writing—review and editing, I.K., L.B. and A.G.; visualization, I.T.P.; supervision, L.B.; project administration, A.G.; funding acquisition, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge the support of this work by the project "Study, Design, Development and Implementation of a Holistic System for Upgrading the Quality of Life and Activity of the Elderly" (MIS 5047294) which is implemented under the Action "Support for Regional Excellence", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

Data Availability Statement: The above research is based on publicly available datasets.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, B.; Zhang, Y. Localization and Tracking of Closely-Spaced Human Targets Based on Infrared Sensors. *Infrared Phys. Technol.* **2022**, *123*, 104176. [CrossRef]
2. Cadena, C.; Carlone, L.; Carrillo, H.; Latif, Y.; Scaramuzza, D.; Neira, J.; Reid, I.; Leonard, J.J. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Trans. Robot.* **2016**, *32*, 1309–1332. [CrossRef]
3. Yin, W.; He, K.; Xu, D.; Luo, Y.; Gong, J. Significant Target Analysis and Detail Preserving Based Infrared and Visible Image Fusion. *Infrared Phys. Technol.* **2022**, *121*, 104041. [CrossRef]
4. Yu, D.; Lin, S.; Lu, X.; Wang, B.; Li, D.; Wang, Y. A Multi-Band Image Synchronous Fusion Method Based on Saliency. *Infrared Phys. Technol.* **2022**, *127*, 104466. [CrossRef]
5. Masone, C.; Caputo, B. A Survey on Deep Visual Place Recognition. *IEEE Access* **2021**, *9*, 19516–19547. [CrossRef]
6. Lowry, S.; Sunderhauf, N.; Newman, P.; Leonard, J.J.; Cox, D.; Corke, P.; Milford, M.J. Visual Place Recognition: A Survey. *IEEE Trans. Robot.* **2016**, *32*, 1–19. [CrossRef]
7. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. The Revisiting Problem in Simultaneous Localization and Mapping: A Survey on Visual Loop Closure Detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 1–25. [CrossRef]
8. Olid, D.; Fácil, J.M.; Civera, J. Single-View Place Recognition under Seasonal Changes. In Proceedings of the 2018 IEEE International Conference on Intelligent Robots and Systems, 10th Planning, Perception and Navigation for Intelligent Vehicles Workshop, Madrid, Spain, 1–5 October 2018.
9. Osman, H.; Darwish, N.; Bayoumi, A. PlaceNet: A Multi-Scale Semantic-Aware Model for Visual Loop Closure Detection. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105797. [CrossRef]
10. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1 Year, 1000 Km: The Oxford RobotCar Dataset. *Int. J. Robot. Res.* **2017**, *36*, 3–15. [CrossRef]
11. Milford, M.J.; Wyeth, G.F. SeqSLAM: Visual Route-Based Navigation for Sunny Summer Days and Stormy Winter Nights. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, St. Paul, MI, USA, 14–18 May 2012; pp. 1643–1649. [CrossRef]
12. Sunderhauf, N.; Neubert, P.; Protzel, P. Are We There Yet? Challenging SeqSLAM on a 3000 Km Journey Across All Four Seasons. In Proceedings of the ICRA 2013 Workshop on Long-Term Autonomy, Karlsruhe, Germany, 6–10 May 2013; p. 3.
13. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
14. Galvez-López, D.; Tardos, J.D. Bags of Binary Words for Fast Place Recognition in Image Sequences. *IEEE Trans. Robot.* **2012**, *28*, 1188–1197. [CrossRef]

15. Sivic, J.; Zisserman, A. Video Google: A Text Retrieval Approach to Object Matching in Videos. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 2, pp. 1470–1477. [CrossRef]
16. Sünderhauf, N.; Shirazi, S.; Jacobson, A.; Pepperell, E.; Dayoub, F.; Upcroft, B.; Milford, M. Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free. In Proceedings of the Robotics: Science and Systems XI, Rome, Italy, 13–17 July 2015; pp. 1–10. [CrossRef]
17. Zaffar, M.; Ehsan, S.; Milford, M.; McDonald-Maier, K. CoHOG: A Light-Weight, Compute-Efficient, and Training-Free Visual Place Recognition Technique for Changing Environments. *IEEE Robot. Autom. Lett.* **2020**, *5*, 1835–1842. [CrossRef]
18. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5297–5307.
19. Gao, P.; Zhang, H. Long-Term Place Recognition through Worst-case Graph Matching to Integrate Landmark Appearances and Spatial Relationships. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation, Paris, France, 31 May–31 August 2020; pp. 1070–1076. [CrossRef]
20. Khaliq, A.; Ehsan, S.; Chen, Z.; Milford, M.; McDonald-Maier, K. A Holistic Visual Place Recognition Approach Using Lightweight CNNs for Significant ViewPoint and Appearance Changes. *IEEE Trans. Robot.* **2020**, *36*, 561–569. [CrossRef]
21. Papapetros, I.T.; Balaska, V.; Gasteratos, A. Visual Loop-Closure Detection via Prominent Feature Tracking. *J. Intell. Robot. Syst.* **2022**, *104*, 54. [CrossRef]
22. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. Modest-Vocabulary Loop-Closure Detection with Incremental Bag of Tracked Words. *Robot. Auton. Syst.* **2021**, *141*, 103782. [CrossRef]
23. Bampis, L.; Amanatiadis, A.; Gasteratos, A. Fast Loop-Closure Detection Using Visual-Word-Vectors from Image Sequences. *Int. J. Robot. Res.* **2018**, *37*, 62–82. [CrossRef]
24. Diamantas, S.; Dasgupta, P. Optical Flow-Based Place Recognition: Bridging the Gap Between Simulation and Real-World Experiments. In Proceedings of the 2022 IEEE 12th Annual Computing and Communication Workshop and Conference, Virtual, 26–29 January 2022; pp. 619–624. [CrossRef]
25. Garg, S.; Harwood, B.; Anand, G.; Milford, M. Delta Descriptors: Change-Based Place Representation for Robust Visual Localization. *IEEE Robot. Autom. Lett.* **2020**, *5*, 5120–5127. [CrossRef]
26. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. Probabilistic Appearance-Based Place Recognition through Bag of Tracked Words. *IEEE Robot. Autom. Lett.* **2019**, *4*, 1737–1744. [CrossRef]
27. Bampis, L.; Gasteratos, A. Sequence-Based Visual Place Recognition: A Scale-Space Approach for Boundary Detection. *Auton. Robot.* **2021**, *45*, 505–518. [CrossRef]
28. Kreyszig, E. Data Analysis. Probability Theory. In *Advanced Engineering Mathematics*, 10th ed.; Wiley: Hoboken, NJ, USA, 2011; p. 1014.
29. Barnes, D.; Gadd, M.; Murcutt, P.; Newman, P.; Posner, I. The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation, Paris, France, 31 May–31 August 2020; pp. 6433–6438. [CrossRef]
30. Pronobis, A.; Caputo, B. COLD: The CoSy Localization Database. *Int. J. Robot. Res.* **2009**, *28*, 588–594. [CrossRef]
31. Buckland, M.; Gey, F. The Relationship between Recall and Precision. *J. Am. Soc. Inf. Sci.* **1994**, *45*, 12–19. [CrossRef]
32. Hausler, S.; Garg, S.; Xu, M.; Milford, M.; Fischer, T. 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, Nashville, TN, USA, 20–25 June 2021; pp. 14141–14152.
33. Khaliq, A.; Milford, M.; Garg, S. MultiRes-NetVLAD: Augmenting Place Recognition Training with Low-Resolution Imagery. *IEEE Robot. Autom. Lett.* **2022**, *7*, 3882–3889. [CrossRef]
34. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893. [CrossRef]
35. Lowry, S.; Andreasson, H. Lightweight, Viewpoint-Invariant Visual Place Recognition in Changing Environments. *IEEE Robot. Autom. Lett.* **2018**, *3*, 957–964. [CrossRef]
36. Han, F.; Yang, X.; Deng, Y.; Rentschler, M.; Yang, D.; Zhang, H. SRAL: Shared Representative Appearance Learning for Long-Term Visual Place Recognition. *IEEE Robot. Autom. Lett.* **2017**, *2*, 1172–1179. [CrossRef]