robotics

MDPI

Article

Place Recognition with Memorable and Stable Cues for Loop Closure of Visual SLAM Systems †

Rafiqul Islam * and Habibullah Habibullah

UniSA STEM, University of South Australia, Mawson Lakes, SA 5095, Australia

* Correspondence: md_rafiqul.islam@mymail.unisa.edu.au

† This paper is an extended version of our paper published in Islam, R.; H. Habibullah, H. A Semantically Aware Place Recognition System for Loop Closure of a Visual SLAM System. In Proceedings of the 2021 4th International Conference on Mechatronics, Robotics and Automation (ICMRA), Zhanjiang, China, 22–24 October 2021.

Abstract: Visual Place Recognition (VPR) is a fundamental yet challenging task in Visual Simultaneous Localization and Mapping (V-SLAM) problems. The VPR works as a subsystem of the V-SLAM. VPR is the task of retrieving images upon revisiting the same place in different conditions. The problem is even more difficult for agricultural and all-terrain autonomous mobile robots that work in different scenarios and weather conditions. Over the last few years, many state-of-the-art methods have been proposed to solve the limitations of existing VPR techniques. VPR using bag-of-words obtained from local features works well for a large-scale image retrieval problem. However, the aggregation of local features arbitrarily produces a large bag-of-words vector database, limits the capability of efficient feature learning, and aggregation and querying of candidate images. Moreover, aggregating arbitrary features is inefficient as not all local features equally contribute to long-term place recognition tasks. Therefore, a novel VPR architecture is proposed suitable for efficient place recognition with semantically meaningful local features and their 3D geometrical verifications. The proposed end-to-end architecture is fueled by a deep neural network, a bag-of-words database, and 3D geometrical verification for place recognition. This method is aware of meaningful and informative features of images for better scene understanding. Later, 3D geometrical information from the corresponding meaningful features is computed and utilised for verifying correct place recognition. The proposed method is tested on four well-known public datasets, and Micro Aerial Vehicle (MAV) recorded dataset for experimental validation from Victoria Park, Adelaide, Australia. The extensive experimental results considering standard evaluation metrics for VPR show that the proposed method produces superior performance than the available state-of-the-art methods.

Keywords: visual place recognition; loop detection; loop-closure; image retrieval

Citation: Islam, R.; Habibullah, H. Place Recognition with Memorable and Stable Cues for Loop Closure of Visual SLAM Systems. *Robotics* **2022**, *11*, 142. https://doi.org/10.3390/robotics11060142

Academic Editor: Rui P. Rocha

Received: 11 September 2022
Accepted: 26 November 2022
Published: 4 December 2022

1. Introduction

Visual Place Recognition (VPR) is the task of accurately and efficiently recognising a revisited place using a camera [1]. This task is challenging in many computer vision and robotics applications. In robotics, the VPR system is often considered an essential task for a V-SLAM. In the V-SLAM problems, the VPR system recognises revisited places of the executed trajectory and creates edge constraints between the present and previously visited pose nodes [2]. Those edges are later used by the Pose Graph Optimization (PGO) algorithm to refine the estimated trajectory and construct an accurate map of an environment [3]. Moreover, the place recognition method helps the V-SLAM system to re-localize or re-initialize efficiently in an already mapped area in case of tracking loss. In autonomous navigation, tracking loss occurs for many reasons, i.e., camera view occlusion, sudden motions, motion blur, system reboot, etc. Therefore, an optimal VPR system is crucial for mobile robotics and autonomous navigation system.

Many state-of-the-art methods have been presented in the last decade to improve VPR performance. The proposed methods from the literature can be subdivided into three main streams: map-to-map, image-to-map, and image-to-image (i2i) matching [2,4]. Map-to-map matching refers to recognising a place on a map by comparing it to a map from the current frame. The image-to-map matching first creates a map of the area using images. It then looks for patterns in the query image that match the patterns in the map. Finally, it uses these patterns to recognise the place on the map. On the other hand, the i2i methods build a database from the image stream collected online by the camera sensor so that the similar one can be retrieved when a similar frame comes from the revisited place. Among these three methods, the i2i method is proven to be successful in large-scale VPR [3,5]. However, this technique mainly suffers from perceptual aliasing. Repetitive features from the scene cause perceptual aliasing, i.e., images primarily covered by grasses, leaves, etc. The perceptual aliasing can be improved by selecting distinct regions from an image [6].

This paper proposes a novel VPR technique that reduces non-meaningful features with semantic segmentation and uses 3D geometrical verification for accurate place recognition. Some information carries more weight than others to recognise a place, i.e., buildings, signs, fences, trees, etc. Therefore, utilising contextual information is efficient for accurate place recognition. Additionally, most autonomous mobile robots require semantic information for many other purposes, i.e., lane and dynamic object detection during autonomous navigation. Therefore, the proposed method will fulfill such requirements while reducing additional computational costs to the system. This paper also demonstrates how a modern Single Board Computer (SBC), i.e., Jetson AGX Xavier, can handle a Deep Neural Network (DNN) model in real-time with high accuracy. The performance of the SBC can be seen in Table 1. This work is an extension of our previously published conference paper [7]. The main contributions of this research are as follows:

1. A semantically aware place recognition method is introduced, efficiently reducing the less meaningful feature information in the database.
2. Optimize the overall image retrieval database by removing unnecessary local features that have less contribution to place recognition.
3. 3D spatial information of the landmarks has been used for correct place recognition, which produces the highest precession and recall accuracy.
4. A new benchmarking Micro Aerial Vehicle (MAV) dataset has been introduced for visual place recognition and V-SLAM evaluation.

Table 1. Semantic segmentation processing rate in different datasets using Jetson Xavier.

Dataset	Resolution	Network	Accuracy	Jetson Xavier
Cityscapes	512 × 256	fcn-resnet18	83.3%	480 FPS
Cityscapes	1024 × 521	fcn-resnet18	87.3%	176 FPS
DeepScene	576 × 320	fcn-resnet18	96.4%	360 FPS
Pascal VOC	512 × 400	fcn-resnet18	64.0%	340 FPS

In the rest of the paper, the sections are discussed in the following orders: related work in Section 2, the proposed method in Section 3, experimental result and discussion in Section 4, respectively, and finally ends with a conclusion by Section 5.

2. Related Work

The intensive literature can justify the significance of the place recognition problem and survey reports from the research communities [8–11]. Even though the current state-of-the-art methods present quite promising progress in different aspects of the place recognition problem, there is still a variety of research scopes for further development, i.e., new requirement, efficiency, scalability, feasibility, robustness, and optimisation.

In recent years, many state-of-the-art methods have been proposed for solving place recognition problems [2,12–18]. Along with traditional handcrafted local feature-based

VPR methods, learning-based methods are becoming more popular in terms of robustness and performance. [19,20]. FAB-MAP [5] was one of the most successful local feature-based approaches from the early days where an appearance-based place recognition using visual words was proposed. FAB-MAP could detect loops with a mono-camera as long as 70 to 1000 km trajectories with no false-positive predictions. However, it was less effective when a big image sequence contained very similar structures. To solve this problem, Bag-of-Words (BoW) methods are presented where an accelerated segment test (FAST)+BRIEF features were used to build a bag-of-words vocabulary tree that discretises a binary descriptor space [13,21,22]. Therefore, it speeds up correspondences for geometrical verification. This novel method secured a recall of 81.20% in the Bicocca25B dataset. The execution time for the whole system requires only 22ms/frame, which is very efficient for real-time applications. Despite its fast runtime, it has some limitations in choosing meaningful visual words for efficient place recognition.

Attention-based visual words have been used to tackle the long-term place recognition problem [23–25]. Arandjelovic et al. [26] proposed a method for learning visually discriminant image regions to create a dense and salient scene description. This approach learns stable image regions over a long period and significant perceptual changes, which means it precisely segments challenging areas of an image. Although this method works well in long-term visual recognition, the system's accuracy can be further improved, and the computational cost can be reduced. In another work [27], a natural language generation framework along with Long Short Term Memory (LSTM) was proposed to imitate the process of place understanding in a human-like nature. However, the method did not consider the decision-making procedure for conflicting VPR matches. Moreover, the performance of the technique can be further improved. Mousavian et al. [28] proposed quite a similar method that leverages segmentation to select features from static objects, i.e., building, to improve the accuracy of the bag-of-words technique. The approach works reasonably well in subtle environmental changes but needs to show more robustness in extreme environmental changes.

Deep visual place recognition is now becoming a popular research area as the purely geometry-based concept does not provide rich information of images of a scene [25,29–31]. In the learning-based domain, a Convolutional Neural Network (CNN) is widely used to learn features in general. The first CNN-based place recognition method was proposed in 2014. CNN-based methods have recently been used as robust feature extractors for place recognition in changing environments. It is found that features from early layers indicate robustness against appearance changes. In contrast, later features are more robust against changes in viewpoint and bear more semantic information that can be utilised to narrow down the search area [32–35]. In place recognition, the CNN was first employed when a Fully Connected (FC) layer of pre-trained ImageNet was effectively used for image retrieval problem [36–38]. Later it was found that if the model is explicitly trained for place recognition using triplet loss, a better result can be achieved with FC representations [39,40]. Even though such techniques show to close the gap with the hand-crafted representations from local descriptors, it becomes computationally expensive and does not solve the limitations of FC layers. Moreover, FC representations are limited to requiring large numbers of parameters and fixed input sizes.

3. Proposed Method

The proposed system consists of mainly three subsystems, Meaningful Feature Selection Section 3.1, Topological Database Section 3.2 using those selected meaningful features, and finally, Geometrical Verification Section 3.3 for avoiding wrong place recognition. The objective of the proposed method is to find meaningful features from a camera frame and recognise a place upon revisiting the same place.

3.1. Meaningful Feature Selection

For meaningful feature selection, an image frame has been segmented into 32 categories for contextual information, and details about the categories can be found in the classes of the CamVid dataset [41]. Image regions consisting of static objects, i.e., trees, fences, bridges, column poles, sidewalks, sign symbols, and buildings, are considered contextual regions for the meaning feature selection. SegNet [42] is used as a base model to segment those interesting regions. The encoders are implemented based on the 13 convolutional layers of the VGG-16 network, while the decoders are implemented with the layers in reverse. Monte Carlo samples of the model are utilised to obtain the probabilistic output. The variance of these softmax samples is taken as the model uncertainty for each class. For the observed training data, D with labels L, the posterior distribution for the weights, W can be found for the network as,

$$p(W \mid D, L) \tag{1}$$

The distribution of the weights, W, needs to be calculated as the posterior distribution is not controllable. Variational inference is used to approximate it. This technique helps us to estimate the distribution of the network's weights, $q(W)$, by minimising the Kullback-Leibler K divergence;

$$K(q(W) \mid\mid p(W \mid D, L)). \tag{2}$$

where, the variational distribution $q(W_i)$ for $N \times N$ dimensional layer i, with unit j is defined as:

$$\begin{aligned} b_{i,j} &\sim Bernoulli(p_i)\ for\ j = 1, ..., M_i, \\ W_i &= M_i diag(b_i) \end{aligned} \tag{3}$$

where, the variational parameter is M_i, vector of Bernoulli distributed random variables is b_i, and the dropout probabilities is $P_i = 0.5$.

After computing the contextual regions, the extracted features are considered meaningful from the image frame. Applying this technique, the extracted features become more efficient and more discriminative.

3.2. Topological Database

The proposed VPR database is called a topological database. The topological database is inspired by the term topology, as this database stores the geometrical properties of spatial objects and preserves the properties under a change of visual perspective. The topological dataset stores descriptors, visual words, and corresponding feature points. The feature descriptors and visual worlds are stored similarly to the DBoW2 database structure. However, the coordinates of the feature points are stored in a separate column for corresponding descriptors. These feature points are later used for geometrical verification. The details of database query and geometrical verification have been discussed in Section 3.3. The structure of the vocabulary tree, direct and inverse indexes are shown in Figure 1. It involves two primary elements. Firstly, a vocabulary W composed of training visual words; secondly, a database consisting of visual words for the candidate, $D = d_1, ..., d_N$, where each entry d represents the bag-of-words associated to a sensor reading at a known pose in the current map.

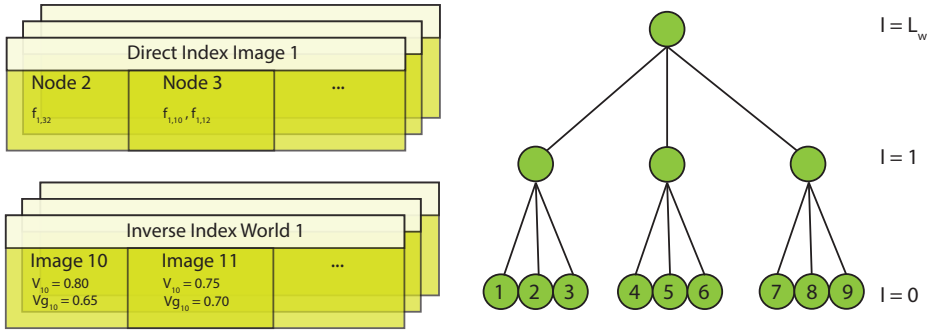


Figure 1. Structure of vocabulary tree and direct and inverse indexes. The inverse index stores the weight of the words in the images. The direct index stores the features of the images and their associated nodes at a certain level of the vocabulary tree.

3.3. Database Query and Geometrical Verification

For place recognition, instantaneous image frames are passed through a meaningful feature selection process as discussed in Section 3.1, which produces local contextual features. Therefore, the local features are queried to retrieve the image index using the potential match similar to the DBoW2 database query. Before considering the correct place recognition, the geometrical verification step is processed. The geometrical verification processes two actions. Firstly, it selects feature points from the active frame and the database. Secondly, it computes similar triangles using the selected features for correct place recognition. These two processes have been further discussed as follows:

3.3.1. Points Picking Policy

The points-picking policy starts with a user-defined input that takes integer numbers. The integer number represents how many triangles to generate for the geometrical verification. Then the features from the current frame are queried in the database to retrieve the frame index and corresponding matched features from the database. Therefore, the Lowe's ratio test is taken to sort out the features based on a good score [43]. From the list of good features, every three best-scored features are chosen for generating a triangle using their corresponding points stored in the database. As previously mentioned, the number of triangles to consider for the geometrical verification is based on user input. The computation of triangles in 3D space and final visual place recognition has been discussed in Section 3.3.2.

3.3.2. Computing Similar Triangle

Similar triangles are triangles that are geometrically identical, and the corresponding sides of similar triangles are always the same. Therefore, the concept of similar triangles has been used for geometrical verification of the proposed VPR. For geometrical verification, the previously chosen meaningful features are fed to PnP [44] method to generate the 3D points. Therefore, similar triangles are computed, one from an instantaneous camera frame and the other from the database, to test whether a scene of an instantaneous frame is similar to the index in the database. Figure 2 shows the process of geometrical verification for correct place recognition. The PNP method doesn't produce a very similar position every time, as it is affected by the camera sensor noise and illumination changes. But it is empirically found that the threshold tolerance works well within less than 2%. The concept of calculating 3D points has been shown in Figure 3. Each side of the triangle is calculated as follows:

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_2 - Z_1)^2} \tag{4}$$

where d = distance and (X, Y, Z) are the elements of Cartesian coordinate system.

Finally, the computed triangles from the current frame and the database are compared following the similar triangle's rules. The proposed system considers this condition a positive visual place recognition if it produces a positive outcome.

Figure 2. Process of geometrical verification for a correct place recognition demonstrated using images from UniSA, Mawson Lakes campus in front of M-Building. In each image, a triangle is constructed using 3D information of the best matching correspondences and compared for correct place recognition.

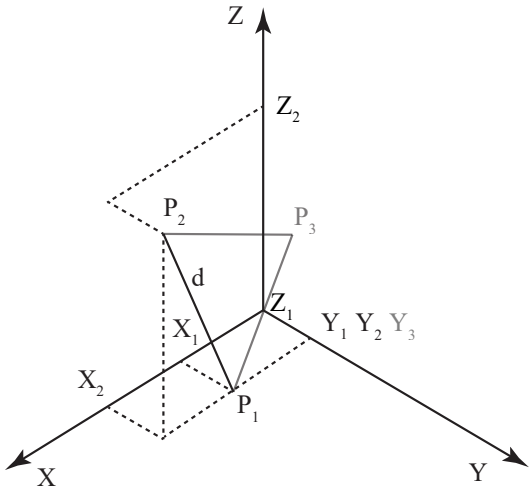


Figure 3. Computing distance of two points in 3D space.

4. Experimental Evaluation

Experimental setup, standard evaluation metrics, and comparative experimental results have been presented in this section. Several experiments have been carried out, and the proposed method has been compared with other state-of-the-art methods for place recognition evaluation. The subsections are presented in the following orders, experimental setup in Section 4.1, evaluation metrics in Section 4.2, and finally, experimental results in Section 4.3.

4.1. Experimental Setup

For the experiment, four standard datasets and our Micro Aerial Vehicle (MAV) recorded dataset were used to evaluate the proposed method with baseline methods. Those datasets are namely Bicocca 2009-02-25b [45], New College [46], Malaga 2009 Parking 6L [47], City Centre [5], and our Victoria Park dataset. The hardware setup for the Victoria Park dataset has been shown in Figure 4. A brief description of the used datasets has been presented in Table 2. All the experiments have been conducted on a real-time single-board computer (NVIDIA's Jetson AGX Xavier 32GB). The specifications of the used hardware are summarised in Table 3.

4.2. Evaluation Metrics

Precision and recall are the two standard evaluation metrics for place recognition approaches. Precision is defined as the proportion of the selected matches that are true positive.

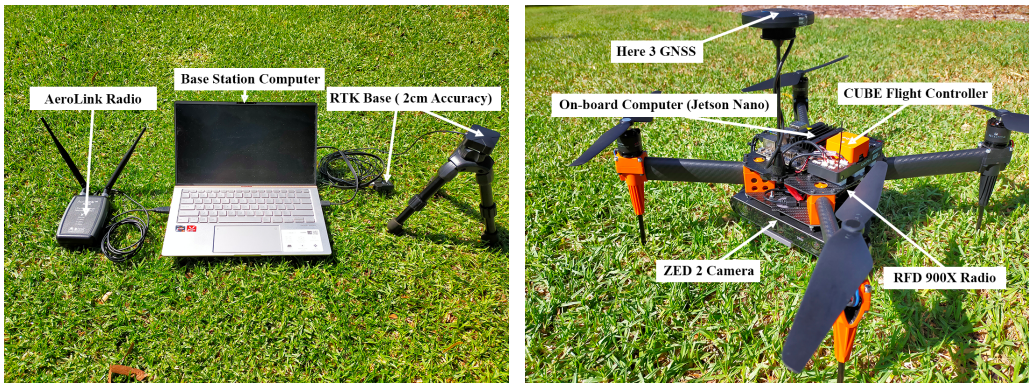
$$precision = \frac{T_p}{T_p + F_p} \tag{5}$$

The recall is the proportion of true positives to the total number of actual matches.

$$recall = \frac{T_p}{T_p + F_n} \tag{6}$$

where T_p = The correct place recognition F_p = the incorrect place recognition F_n = system thought it was a correct place recognition while it was incorrect.

A perfect system should achieve precision and recall of 100%. Precision and recall are usually illustrated using a precision-recall curve, which plots recall against precision for a range of confidence scores. Until now, avoiding false positive matches for place recognition is recommended, as accepting false place recognition for optimising the map can cause catastrophic failure. Therefore, the key metric to measure the place recognition system is recall at 100% precision [8].



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

Figure 4. The base station and Micro Aerial Vehicle (MAV) setup for the recorded Victoria Park dataset: (**a**) shows the base station consists of Aerolink radio telemetry, RTK base having 2cm of accuracy. A base station computer; (**b**) shows the setup of the MAV consisting of Here 3 Global Navigation Satellite System (GNSS), Jetson nano onboard computer, CUBE flight controller, ZED 2 Stereo Camera and RFD 900X long-range radio telemetry. Both the base station and the MAV are wirelessly linked by long-range radio telemetry.

Table 2. Properties of the Benchmarking datasets.

Dataset	**Description**	**Image Size (px × px)**
New College [46]	Outdoors, Dynamic Objects, Frontal Camera	512 × 384 @ 20 Hz
Bicocca [45]	Indoors, Static Objects, Frontal Camera	640 × 480
City Centre [5]	Urban Area, Dynamic Objects, Lateral Camera	640 × 480
Malaga [47]	Outdoors, Dynamic Objects, Lateral Camera	1024 × 768
Victoria Park, Adelaide, Australia [Ours]	Outdoors, Dynamic Objects, Frontal Stereo Camera	1024 × 720 @ 30 Hz

Table 3. Specifications of the experimental hardwires.

JETSON AGX XAVIER		ZED 2 CAMERA	
GPU	512-core Volta GPU with Tensor Cores	Depth FPS	Up to 100 Hz
CPU	8-core ARM v8.2 64-bit CPU, 8 MB L2 + 4 MB L3	Depth Range	0.2–20 m (0.65 to 65 ft)
Memory	32 GB 256-Bit LPDDR4× | 137 GB/s	Sensors	Accelerator, Gyroscope, Barometer, Magnetometer, Temperature Sensor
Storage	32 GB eMMC 5.1	Lens	Wide-angle with optically corrected distortion
DL Accelerator	(2×) NVDLA Engines	Field of View	110° (H) × 70° (V) × 120° (D) max.
Vision Accelerator	7-way VLIW Vision Processor	Aperture	$f/1.8$
Encoder/Decoder	(2×) 4Kp60 | HEVC/(2×) 4Kp60 | 12-Bit Support	Sensor Resolution	Dual 4M pixels sensors with 2-micron pixels

4.3. Experimental Results

Victoria Park Dataset is a stereo camera recorded dataset using the UniSA MAV. This dataset has been recorded from different geo-positions and altitudes, as shown in Figure 5. The recorded sequence consists of many loops from different altitudes and orientations, making the dataset more realistic for evaluating a vision-based robotic system in 3D space. The dataset consists of stereo camera sequence, GPS, Baro, IMU, Compass, and other sensor data. This dataset has different challenges, such that patterns are mostly repetitive with grass, trees, and distant house-like objects. Moreover, the MAV pose changes rapidly, causing the different angles of view for the same scene, which puts the place recognition system into a challenge. The data was recorded just before sunset, so various light conditions combine sunny and shaded areas in the image sequence. As it is a MAV recorded dataset, half of each frame is mainly covered by the sky. The rest of the frame is covered by primarily repetitive patterns as there are many loops in the dataset from different directions that a visual place recognition system can utilise.

Figure 6 shows the recorded dataset's latitude, longitude, and altitude. The background colour of each plot shows different flight modes, as shown in the legends. A few flight modes have been used for the dataset recording, Stabilize, RTL, Loiter, Brake, Drift, Alt Hold (altitude hold), and Land. The Stabilize mode attempts to self-level the roll and pitch axis of the MAV during the flight. The RTL mode (Return To Launch mode) navigates the MAV from its current position to the home position. The Loiter mode automatically attempts to maintain the current Global Positioning System (GPS) location, heading and altitude. The Brake mode stops the MAV as soon as possible once it is triggered. The dataset is mostly recorded with Loiter flight mode. It is shown in Figure 5a how the geo-position changed when the dataset was recorded. The varied geo-positions and altitude affect the performance of the visual place recognition systems which has been discussed in the later part of the experimental results.

An excellent visual place recognition system should recognise a place from a different angle of view for the same scene. A different angle of view can be correlated with camera position and orientation as any 3D world points can be viewed by the appropriate translation and orientation of the camera facing to that point. Hence, our MAV-recorded Victoria Park dataset has been recorded with varied orientations and positions of a scene to challenge any VPR systems. Alongside the stereo camera sequence, Inertial Measurement Unit (IMU) data from three multiple sensors have also been captured in the Victoria Park dataset. Later Extended Kalman Filter (EKF) is used to fuse all the data from different parallel sensors to get a better pose estimation. In Figure 7, the pose data have been presented in Euler angles (Roll, Pitch, and Yaw), where the Euler angle (in degree) is plotted in the y-axis for the timestamp in the x-axis. It is shown later that the visual place recognition systems produce poor results when the pose of the MAV significantly changes for the same scene. All the sensory data of the Victoria Park dataset are synced with a timestamp.

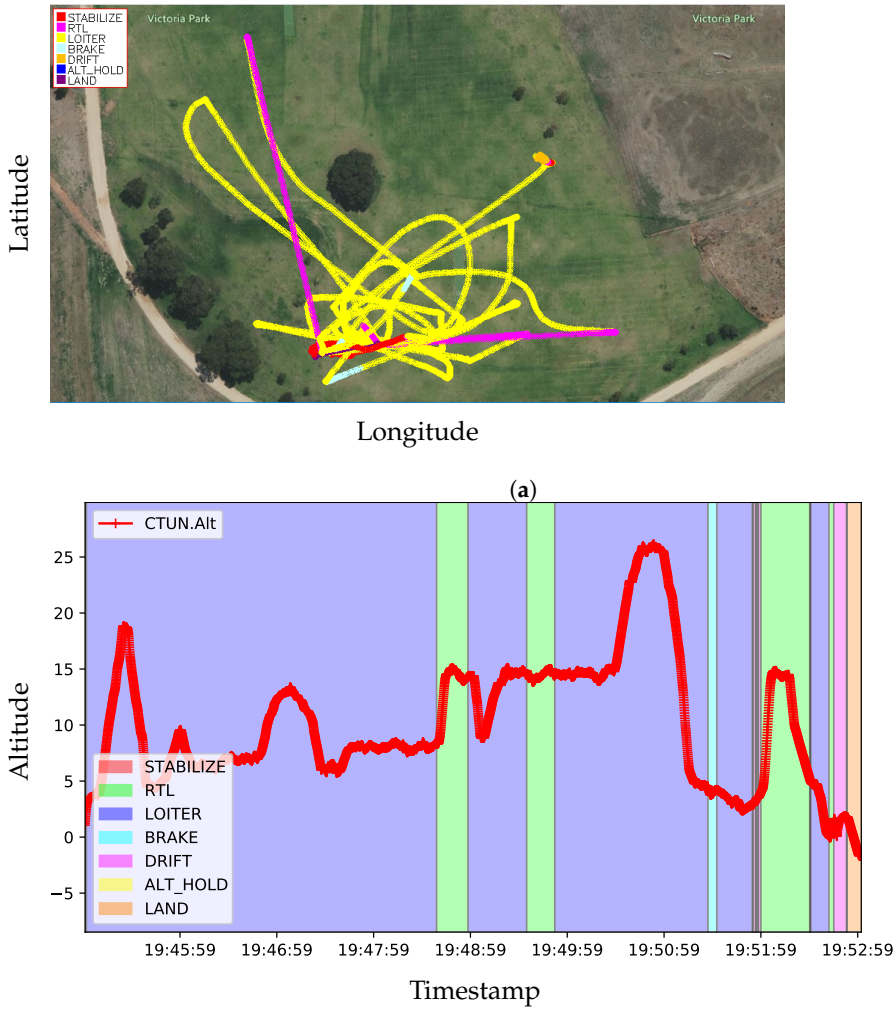


Figure 5. Trajectory and altitude of the Victoria Park dataset: (**a**) shows geo-position (latitude, longitude) of the Victoria Park dataset (Sequence VP-01); (**b**) shows the altitude (m) with respect to time while recording the data, and the background of the plot shows different flight modes during that timestamp.

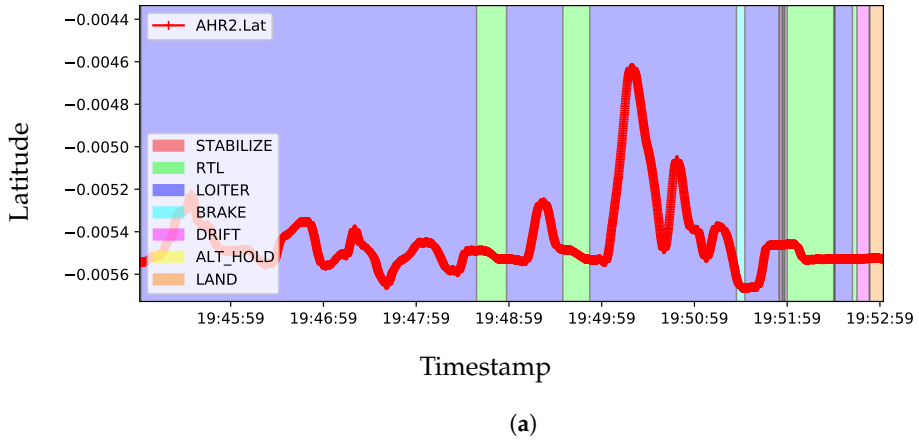


Figure 6. *Cont.*

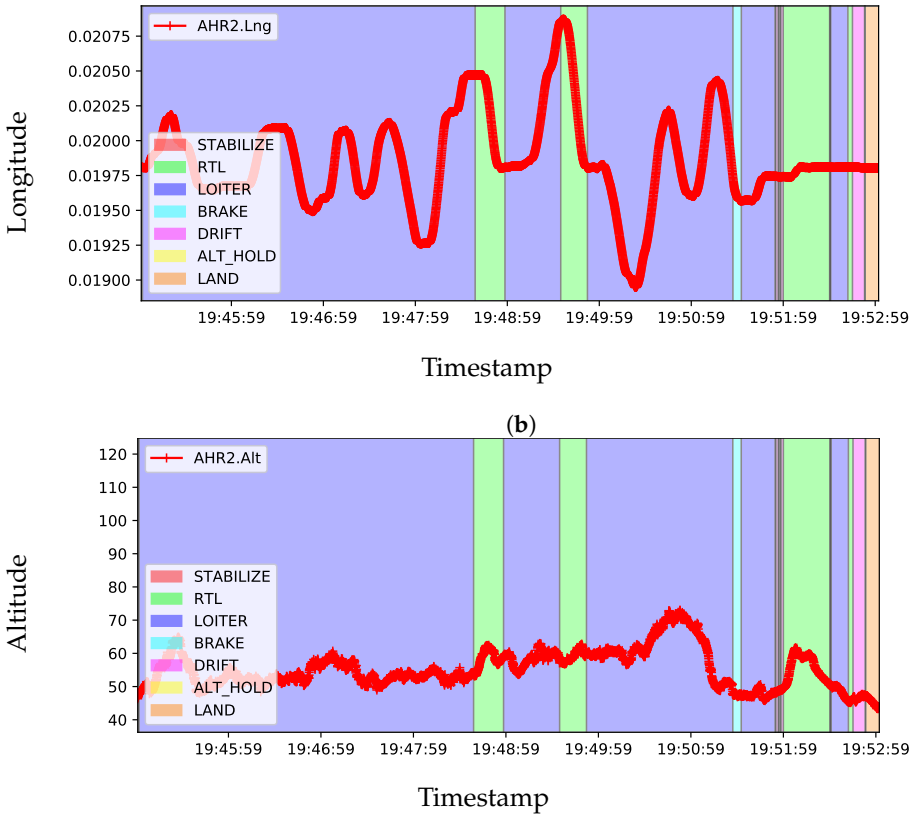


Figure 6. Latitude (**a**), longitude (**b**) and altitude (**c**) of the Victoria Park dataset.

Figure 8 shows the comparison of feature matching from different viewpoints, meaning having different geo-position and orientations with respect to previous frames. The changes in orientation and position of the presented frames are shown in Figures 6 and 7. The matching lines are presented in diverging colourmaps, where the green matching lines represent the more confidence level of the corresponding points and blue represents the lower confidence level. In Figure 8, the left column shows the correspondences using the method by Sarlin [48], where it can be seen the miss matches of corresponding points in the Victoria Park dataset for its highly repetitive patterns and illumination changes. Conversely, the proposed method produces more accurate and stable matches with significant viewpoint and illumination changes. Most importantly, it avoids less meaningful features like grasses and tree shadows. For this challenging dataset, the proposed method finds a good number of correspondences with high confidence to ensure a correct place recognition.

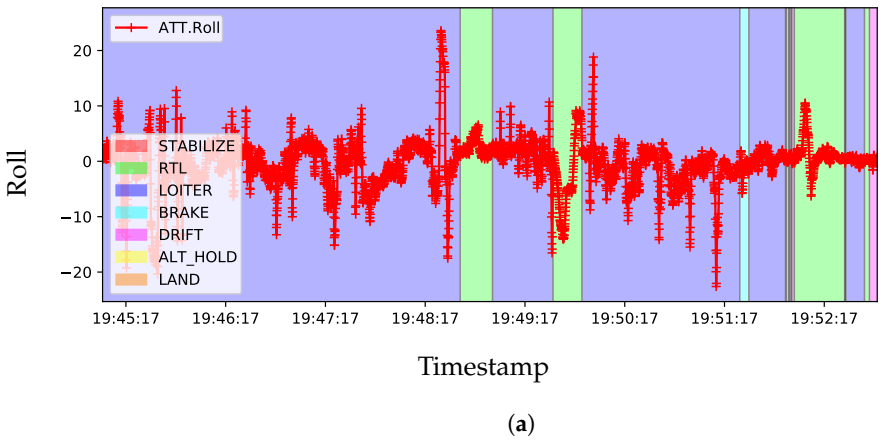


Figure 7. *Cont.*

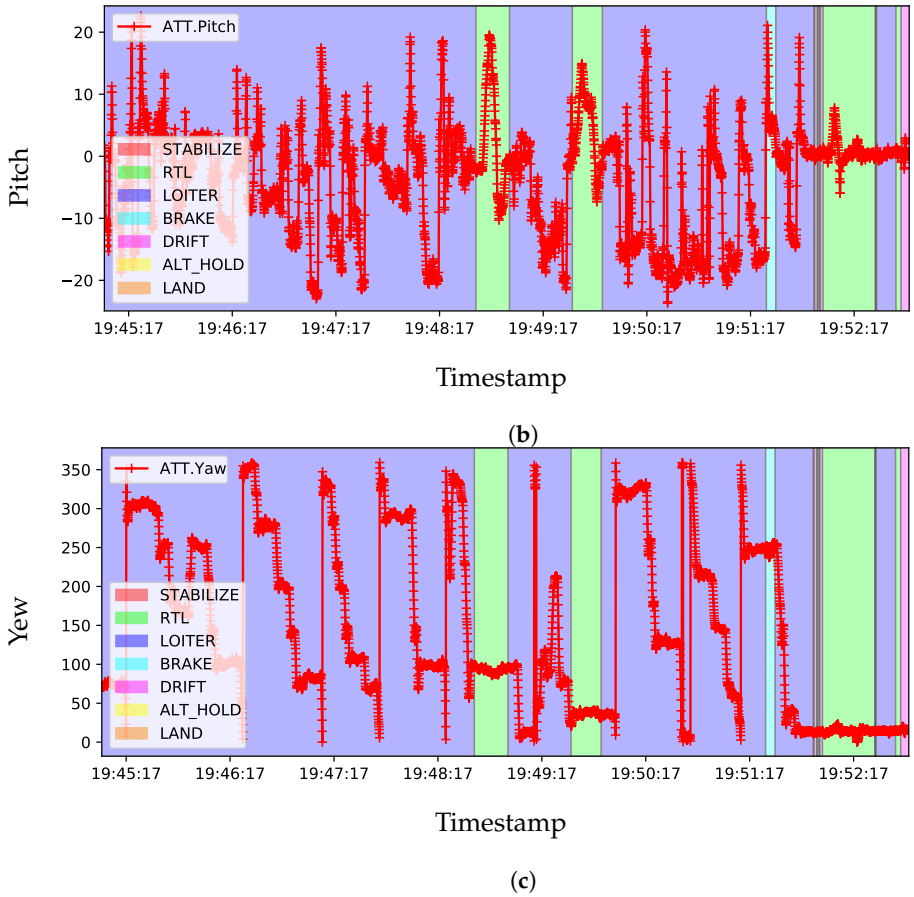


Figure 7. Pose of MAV in Euler angle (deg). Roll (**a**), pitch (**b**) and yew (**c**) for the Victoria Park dataset.

Figure 9 shows the similarity matrix for different methods on the New College dataset [46]. The similarity data have been presented in heatmaps, and each subfigure shows the heatmap of the 4024 images. Each pixel of the subfigure shows the similarity score of that candidate compared with the entries in the database. The lighter the image's pixel intensity, the more similar the image in the database. Figure 9a shows the ground truth for the New College dataset. It can be seen that the background colour is very plain as the intensity of the comparison is zero. Therefore, the whole plain background has the same similarity score. The dark pixel shows the overlapping high similarity score with the entries. Light background with a highly dark foreground represents a desirable higher singular score. When the background and foreground colours are blended, the place recognition system disregards the place recognition. If Figure 9b,c are compared considering the above measures, it can be seen that the proposed method produces the best similarity matrix, which is quite close to the ground truth similarity matrix.

Figure 10a illustrates the relationship between precision and recall for a generic system. The recall is presented on the X-axis and precision on the Y-axis with a 0.1 step size on both axes. The dot point in the curve represents the precision and recall rate for a certain detection threshold. Less precision means more false-positive predictions by the system. In other words, the system is prone to false predictions if the curve moves downward. Recall, on the other hand, reduces if the false-negative prediction increases. Therefore, if the curve tends to the Y-axis, the system is prone to more false predictions that are correct in the ground truth.

Figure 10b shows the precision–recall curve in the New College [46] dataset. The curves have been obtained for different detection thresholds, while other methods achieve a reasonably good result, the proposed method secures as high recall as 75% at 100% precision. As the proposed method utilises 3D geometrical information of the environment,

achieving a minimal and sensitive loop-closure detection threshold is possible, which is impossible using other existing techniques.

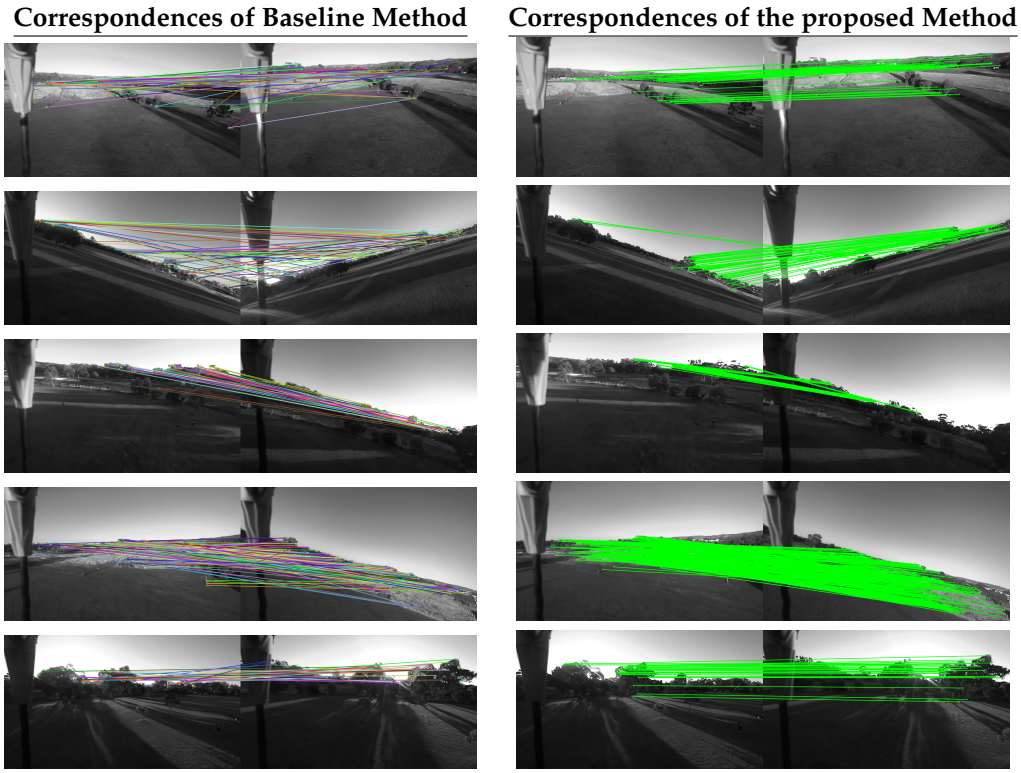
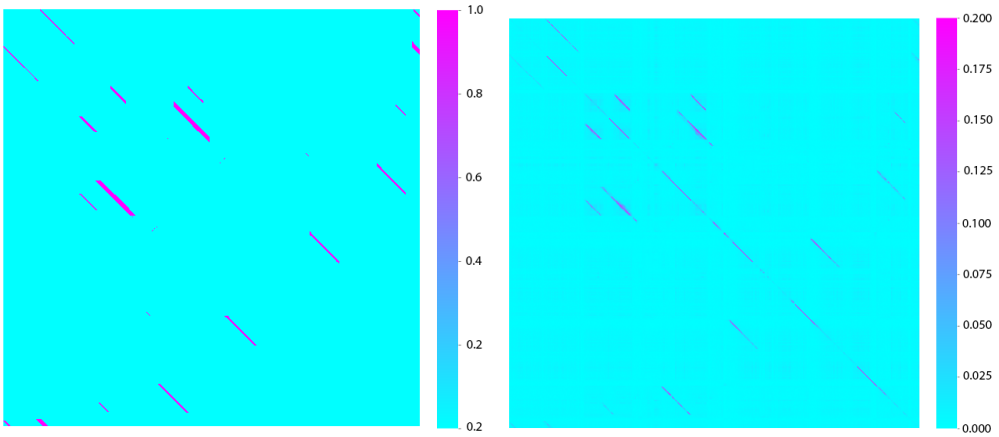
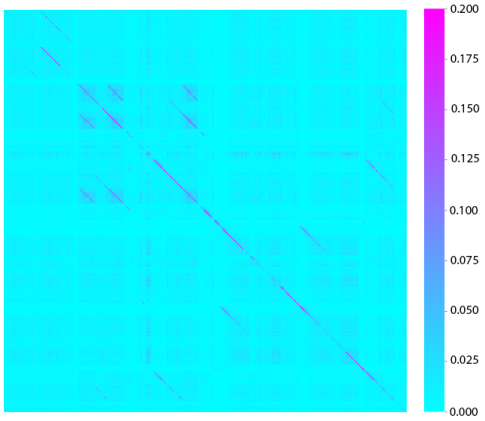


Figure 8. Comparison of corresponding points in the Victoria Park dataset, where more green correspondences mean more accurate place recognition. Images in the left column show corresponding points by experimenting with the VPR method of Sarlin (2020). Images in the right column show corresponding points by the proposed VPR method. The result by the baseline method shows fewer green correspondences; on the other hand, the result by the proposed method shows more green correspondences. The proposed method consistently estimates more correct matches with large viewpoints and illumination changes.



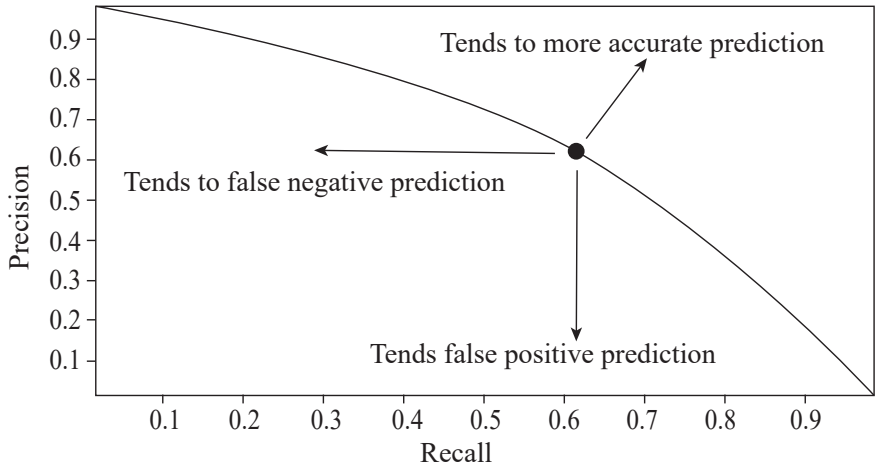
(**a**) (**b**)

Figure 9. *Cont.*

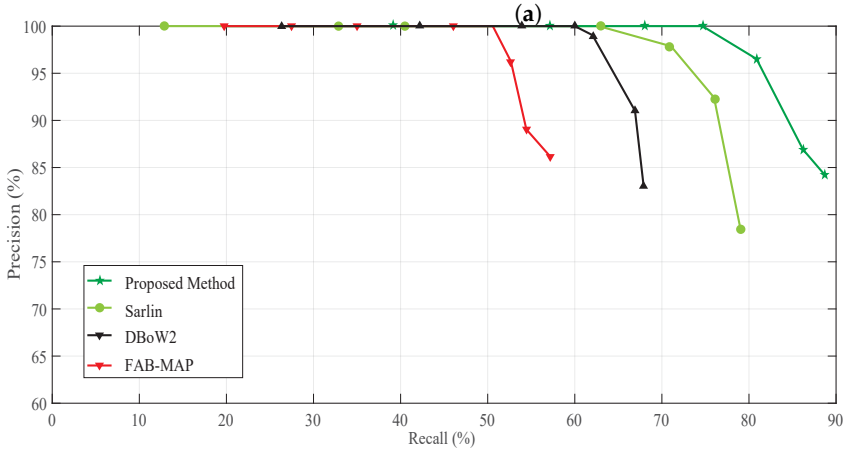


(**c**)

Figure 9. Similarity Matrix comparison to New College dataset. (**a**) shows ground truth, (**b**) shows the proposed method's result, and (**c**) shows the result of DBoW Lopez (2012).



(**a**)



(**b**)

Figure 10. Precision and recall accuracy demonstrated in the curve: (**a**) performance relationship of place recognition in the precision–recall curve; (**b**) precision–recall curve in the New College dataset Smith (2009).

Comparative results are summarised in Table 4, where the proposed method has been compared with other state-of-the-art methods on the same datasets. The compared baseline methods are HF-Net [49], DBoW2 [13], and FAB-MAP [5]. The performance of each of the baseline methods has been obtained from experimenting with their open-

source implementations. HF-Net by Sarlin [49] works relatively well in dynamic lighting conditions and achieves better precision and recall accuracy than DBoW2 [13] and FAB-MAP [5]. However, the network needs intensive GPU computation costs, and it could reach up to 8 FPS with NVIDIA RTX 3060 GPU. Therefore, the method is not suitable for low-powered robotic applications. On the other hand, the proposed method produces better precision and recall accuracy in most of the datasets except the Malaga [47] dataset, where Sarlin [49] achieves a slightly better result than the proposed method. Most importantly, the proposed method obtains significantly better precision and recall accuracy while using less computational resources. The proposed method obtained 65% recall at 100% precision in the City Centre [5], which is considered one of the most complex datasets in this domain.

Table 4. Comparative results on Precision and Recall.

Dataset	Methods	Precision (%)	Recall (%)
New College	Proposed	100	75.80
	Sarlin [49]	100	62.25
	DBoW2 [13]	100	60.12
	FAB-MAP [5]	100	52.54
Malaga6L	Proposed	100	77.45
	Sarlin [49]	100	78.42
	DBoW2 [13]	100	72.56
	FAB-MAP [5]	100	65.87
Bicocca25b	Proposed	100	85.32
	Sarlin [49]	100	80.20
	DBoW2 [13]	100	56.52
	FAB-MAP [5]	100	N/A
City Center	Proposed	100	65.45
	Sarlin [49]	100	45.15
	DBoW2 [13]	100	32.73
	FAB-MAP [5]	100	36.64
Victoria Park	Proposed	100	72.23
	Sarlin [49]	100	53.50
	DBoW2 [13]	100	42.25
	FAB-MAP [5]	100	38.70

The execution time, another important evolution matrix, has been presented in Table 5. Mean, standard deviation, minimum and maximum time taken by the proposed method have been summarised based on different tasks in the program pipelines. The attention network took 2.85 ms of time for each frame computation and, at most, 5.65 ms for a complex operation. The Jetson AGX Xavier has a 7.2 compute-capable GPU, specifically designed for machine learning, deep learning, and mobile robotics. Therefore, the inference mean-time for each frame reduces significantly as efficiency as approximately 2.8 ms. The sum of the whole program execution time is relatively small, around 23.63 ms; therefore, the system can be run on a real-time system with approximately FPS 48.

Table 5. Execution time in New College dataset.

Tasks	Execution Time (ms)			
	Mean	Std	Min	Max
Attention Network	2.85	0.45	1.52	5.65
Feature Extraction	10.81	6.44	7.00	56.00
Bag of Words	7.47	3.54	3.00	25.00
Geometrical Verification	2.50	2.41	1.48	10.50
Total Time	23.63	12.84	13.00	96.00

5. Conclusions

This paper proposes a VPR architecture using semantically and spatially meaningful information from images. Additionally, an optimal place recognition data structure has been designed to aggregate semantical and spatial information for efficient scene retrieval. The semantically meaningful information has been computed using convolutional neural network architecture, and spatially meaningful information has been calculated using the geometrical relationship of 3D landmarks and their corresponding best-scored descriptors to verify and mitigate the effect of false place recognition predictions. The proposed method has been tested on popular standard datasets, and the Micro Aerial Vehicle (MAV)-recorded Victoria Park dataset for the performance evaluation. The proposed method outperforms other state-of-the-art local-feature-based and deep-learning-based methods on different datasets and challenging scenarios while significantly improving the performance with a unique viewpoint and appearance variations. The experimental results are presented using standard evaluation metrics for a better comparison with other state-of-the-art methods. The results show that the proposed method is significantly more accurate in the Victoria Park dataset than the other state-of-the-art methods. For the feasibility of using the system in a real-time application, it has been tested on an NVIDIA Jetson AGX Xavier AI computer with built-in 7.2 compute capable CUDA-Enabled GPU, specially designed for autonomous mobile devices robotics. A possible future direction to extend this work can be introduced by creating an unsupervised scene learning, aggregation, and retrieval with an efficient scene retrieval database.

Author Contributions: Conceptualization, methodology, formal analysis and investigation, writing—original draft preparation: R.I.; review, editing, and supervision: H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Available upon request.

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

Abbreviations

BoW	Bag-of-Words
CPU	Central Processing Unit
CNN	Convolutional Neural Network
DNN	Deep Neural Network
EKF	Extended Kalman Filter
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GPU	Graphics Processing Unit
IMU	Inertial Measurement Unit
i2i	Image-to-Image
LSTM	Long Short Term Memory
MAV	Micro Aerial Vehicle
PGO	Pose Graph Optimization
PnP	Perspective-n-Point
RTK	Real-Time Kinematic
SBC	Single Board Computer
SLAM	Simultaneous Localization and Mapping
UniSA	University of South Australia
UniSA-MLK	University of South Australia—Mawson Lakes
VPR	Visual Place Recognition
V-SLAM	Visual Simultaneous Localization and Mapping

References

1. Zeng, Z.; Zhang, J.; Wang, X.; Chen, Y.; Zhu, C. Place Recognition: An Overview of Vision Perspective. *Appl. Sci.* **2018**, *8*, 2257. [CrossRef]
2. Bampis, L.; Amanatiadis, A.; Gasteratos, A. Encoding the description of image sequences: A two-layered pipeline for loop closure detection. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Republic of Korea, 9–14 October 2016; pp. 4530–4536.
3. Kümmerle, R.; Grisetti, G.; Strasdat, H.; Konolige, K.; Burgard, W. g2o: A general framework for graph optimization. In Proceedings of the IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 3607–3613.
4. Williams, B.; Cummins, M.; Neira, J.; Newman, P.; Reid, I.; Tardós, J.D. A comparison of loop closing techniques in monocular SLAM. *Robot. Auton. Syst.* **2009**, *57*, 1188–1197. [CrossRef]
5. Cummins, M.; Newman, P. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *Int. J. Robot. Res.* **2008**, *27*, 647–665. [CrossRef]
6. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
7. Islam, R.; Habibullah, H. A Semantically Aware Place Recognition System for Loop Closure of a Visual SLAM System. In Proceedings of the 2021 4th International Conference on Mechatronics, Robotics and Automation (ICMRA), Zhanjiang, China, 22–24 October 2021; pp. 117–121.
8. Lowry, S.; Sünderhauf, 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]
9. Torralba, A.; Murphy, K.P.; Freeman, W.T.; Rubin, M.A. Context-based vision system for place and object recognition. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13-16 October 2003; Volume 1, pp. 273–280.
10. Nicosevici, T.; García, R. Automatic Visual Bag-of-Words for Online Robot Navigation and Mapping. *IEEE Trans. Robot.* **2012**, *28*, 886–898. [CrossRef]
11. Lerma, C.D.C.; Gálvez-López, D.; Tardós, J.D.; Neira, J. Robust Place Recognition With Stereo Sequences. *IEEE Trans. Robot.* **2012**, *28*, 871–885.
12. Nistér, D.; Stewénius, H. Scalable Recognition with a Vocabulary Tree. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 2161–2168.
13. Gálvez-López, D.; Tardós, J.D. Bags of Binary Words for Fast Place Recognition in Image Sequences. *IEEE Trans. Robot.* **2012**, *28*, 1188–1197. [CrossRef]
14. Koniusz, P.; Yan, F.; Gosselin, P.H.; Mikolajczyk, K. Higher-Order Occurrence Pooling for Bags-of-Words: Visual Concept Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 313–326. [CrossRef]
15. Keetha, N.V.; Milford, M.; Garg, S. A Hierarchical Dual Model of Environment- and Place-Specific Utility for Visual Place Recognition. *IEEE Robot. Autom. Lett.* **2021**, *6*, 6969–6976. [CrossRef]
16. Bhutta, M.U.M.; Sun, Y.; Lau, D.; Liu, M. Why-So-Deep: Towards Boosting Previously Trained Models for Visual Place Recognition. *IEEE Robot. Autom. Lett.* **2022**, *7*, 1824–1831. [CrossRef]
17. 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]
18. Cai, K.; Wang, B.; Lu, C.X. AutoPlace: Robust Place Recognition with Single-chip Automotive Radar. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 2222–2228.
19. Cai, Y.; Zhao, J.; Cui, J.; Zhang, F.; Ye, C.; Feng, T. Patch-NetVLAD+: Learned patch descriptor and weighted matching strategy for place recognition. In Proceedings of the 2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Bedford, UK, 20–22 September 2022; pp. 1–8.
20. 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 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 14136–14147.
21. Dietsche, A.; Ott, L.; Siegwart, R.Y.; Brockers, R. Visual Loop Closure Detection for a Future Mars Science Helicopter. *IEEE Robot. Autom. Lett.* **2022**, *7*, 12014–12021. [CrossRef]
22. Xin, Z.; Cai, Y.; Lu, T.; Xing, X.; Cai, S.; Zhang, J.; Yang, Y.; Wang, Y. Localizing Discriminative Visual Landmarks for Place Recognition. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 5979–5985.
23. Schönberger, J.L.; Pollefeys, M.; Geiger, A.; Sattler, T. Semantic Visual Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6896–6906.
24. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
25. Masone, C.; Caputo, B. A Survey on Deep Visual Place Recognition. *IEEE Access* **2021**, *9*, 19516–19547. [CrossRef]
26. Naseer, T.; Oliveira, G.L.; Brox, T.; Burgard, W. Semantics-aware visual localization under challenging perceptual conditions. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 2614–2620.

27. Li, P.; Li, X.; Li, X.; Pan, H.; Khyam, M.O.; Noor-A-Rahim, M.; Ge, S.S. Place perception from the fusion of different image representation. *Pattern Recognit.* **2021**, *110*, 107680. [CrossRef]
28. Mousavian, A.; Kosecka, J.; Lien, J.M. Semantically guided location recognition for outdoors scenes. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 4882–4889.
29. Arandjelović, R.; Gronát, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1437–1451. [CrossRef]
30. Li, X.; Li, X.; Khyam, M.O.; Luo, C.; Tan, Y. Visual navigation method for indoor mobile robot based on extended BoW model. *CAAI Trans. Intell. Technol.* **2017**, *2*, 142–147. [CrossRef]
31. Ali-bey, A.; Chaib-draa, B.; Giguère, P. GSV-Cities: Toward Appropriate Supervised Visual Place Recognition. *Neurocomputing* **2022**, *513*, 194–203. [CrossRef]
32. Sünderhauf, N.; Dayoub, F.; Shirazi, S.A.; Upcroft, B.; Milford, M. On the performance of ConvNet features for place recognition. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–3 October 2015; pp. 4297–4304.
33. Zhou, B.; Lapedriza, À.; Xiao, J.; Torralba, A.; Oliva, A. Learning Deep Features for Scene Recognition using Places Database. In Proceedings of the NIPS, Montreal, QC, Canada, 8–13 December 2014.
34. Zaffar, M.; Ehsan, S.; Milford, M.; Flynn, D.; Mcdonald-Maier, K.D. VPR-Bench: An Open-Source Visual Place Recognition Evaluation Framework with Quantifiable Viewpoint and Appearance Change. *Int. J. Comput. Vis.* **2021**, *129*, 2136–2174. [CrossRef]
35. Jiwei, N.; Feng, J.M.; Xue, D.; Feng, P.; Wei, L.; Jun, H.; Cheng, S. A Novel Image Descriptor with Aggregated Semantic Skeleton Representation for Long-term Visual Place Recognition. *arXiv* **2022**, arXiv:abs/2202.03677.
36. Razavian, A.S.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 512–519.
37. Gong, Y.; Wang, L.; Guo, R.; Lazebnik, S. Multi-scale orderless pooling of deep convolutional activation features. In Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 392–407.
38. Liu, Y.; Guo, Y.; Wu, S.; Lew, M.S. Deepindex for accurate and efficient image retrieval. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, 23–26 June 2015; pp. 43–50.
39. Wan, J.; Wang, D.; Hoi, S.C.H.; Wu, P.; Zhu, J.; Zhang, Y.; Li, J. Deep learning for content-based image retrieval: A comprehensive study. In Proceedings of the ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 157–166.
40. Gomez-Ojeda, R.; Lopez-Antequera, M.; Petkov, N.; Gonzalez-Jimenez, J. Training a convolutional neural network for appearance-invariant place recognition. *arXiv* **2015**, arXiv:1505.07428.
41. Brostow, G.J.; Fauqueur, J.; Cipolla, R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognit. Lett.* **2009**, *30*, 88–97. [CrossRef]
42. Kendall, A.; Badrinarayanan, V.; Cipolla, R. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. *arXiv* **2017**, arXiv:1511.02680.
43. LoweDavid, G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
44. Lepetit, V.; Moreno-Noguer, F.; Fua, P. EPnP: An Accurate O(n) Solution to the PnP Problem. *Int. J. Comput. Vis.* **2008**, *81*, 155–166. [CrossRef]
45. Bonarini, A.; Burgard, W.; Fontana, G.; Matteucci, M.; Sorrenti, D.G.; Tardos, J.D. Rawseeds: Robotics advancement through web-publishing of sensorial and elaborated extensive data sets. In Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Beijing, China, 9–15 October 2006; Volume 6, p. 93.
46. Smith, M.; Baldwin, I.; Churchill, W.; Paul, R.; Newman, P. The New College Vision and Laser Data Set. *Int. J. Robot. Res.* **2009**, *28*, 595–599. [CrossRef]
47. Blanco, J.; Moreno, F.; González, J. A collection of outdoor robotic datasets with centimeter-accuracy ground truth. *Auton. Robot.* **2009**, *27*, 327–351. [CrossRef]
48. Sarlin, P.E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning Feature Matching With Graph Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4937–4946.
49. Sarlin, P.E.; Cadena, C.; Siegwart, R.; Dymczyk, M. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12708–12717.