Exploring Performance Bounds of Visual Place Recognition Using Extended Precision

Bruno Ferrarini, Maria Waheed, Sania Waheed, Shoaib Ehsan, Michael J. Milford, and Klaus D. McDonald-Maier

***Abstract*—Recent advances in image description and matching allowed significant improvements in Visual Place Recognition (VPR). The wide variety of methods proposed so far and the increase of the interest in the field have rendered the problem of evaluating VPR methods an important task. As part of the localization process, VPR is a critical stage for many robotic applications and it is expected to perform reliably in any location of the operating environment. To design more reliable and effective localization systems this letter presents a generic evaluation framework based on the new Extended Precision performance metric for VPR. The proposed framework allows assessment of the upper and lower bounds of VPR performance and finds statistically significant performance differences between VPR methods. The proposed evaluation method is used to assess several state-of-the-art techniques with a variety of imaging conditions that an autonomous navigation system commonly encounters on long term runs. The results provide new insights into the behaviour of different VPR methods under varying conditions and help to decide which technique is more appropriate to the nature of the venture or the task assigned to an autonomous robot.**

***Index Terms*—Performance evaluation and benchmarking, visual-based navigation, localization.**

I. Introduction

Visual place recognition (VPR) represents the ability of a robot to decide whether an image shows a previously visited place. VPR is a fundamental task in many endeavours in the field of robotics and hence has been subject to great advancements in recent times in regard to both existing algorithms and new techniques [27]. Often VPR approaches are mutually compared in order to develop a better understanding of the advantages and disadvantages of each technique and attains its full potential during the employment period. Among the state-of-the-art methods some have received limited but prior attention in terms of performance comparison to each other in [56]. VPR techniques are often rated on their performance on different datasets, each having a different intensity of changing variables including illumination [34], [42], presence of dynamic objects [7], [55], viewpoint [19], [28] and seasonal variations [35], [38]. These factors yield changes in the appearance of places, which is the main reason for VPR remains a challenge in autonomous robotic navigation. Though it has been evident through several experiments that each VPR technique might have some perks or an edge when working with a particular dataset and appearance changes [49] but the extent of the critical analysis and comparison among this performance difference still remains an untapped territory.

This letter proposes a new performance metric denoted as Extended Precision (EP) and an evaluation framework which aims to tackle the potentially overlooked features in previous VPR performance comparisons. The evaluation process consists of two phases. The first explores the upper and lower performance bounds of VPR techniques across an environment in order to assess the reliability of the image matching in a changing environment. The second phase uses a statistical approach to identify performance differences between VPR methods. EP is obtained by combining several features of a Precision-Recall Curve into a scalar value which is used in the evaluation framework to measure VPR performance and carry out statistical tests. The proposed framework is then employed to assess several state-of-the-art VPR methods over different datasets, each presenting different types of environmental changes. The results provide new insights into the behaviour of VPR under varying conditions and can give an indication on the more appropriate technique to employ according to the nature of the venture or the task assigned to an autonomous robot.

The rest of the letter is organized as follows. Section II provides an overview of related work. Section III describes the proposed framework and metric. The experimental results are presented and discussed in Section IV. Finally, conclusions are given in Section V.

II. Related Work

Visual place recognition (VPR) is an arduous endeavour in the field of robotic navigation, with the primary goal to accurately recognize a location from visual information. Despite the significant advancements in recent years, VPR still remains a

Manuscript received September 9, 2019; accepted January 6, 2020. Date of publication January 24, 2020; date of current version February 7, 2020. This letter was recommended for publication by Associate Editor J. Guerrero and E. Marchand upon evaluation of the reviewers' comments. This work was supported in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grants EP/R02572X/1 and EP/P017487/1 and in part by RICE Project funded by the National Centre for Nuclear Robotics (NCNR) Flexible Partnership Fund. *(Corresponding author: Bruno Ferrarini.)*

B. Ferrarini, S. Ehsan, and K. D. McDonald-Maier are with CSSE School, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: bferra@essex.ac.uk; sehsan@essex.ac.uk; kdm@essex.ac.uk).

M. Waheed is with COMSATS University, Islamabad 45550, Pakistan (e-mail: maria.waheed97@gmail.com).

S. Waheed is with the National University of Sciences and Technology, Islamabad 45550, Pakistan (e-mail: saniawaheed97@gmail.com).

M. J. Milford is with the Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD 4000, Australia (e-mail: michael.milford@qut.edu.au).

Digital Object Identifier 10.1109/LRA.2020.2969197

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

perfectible task due to the extreme viewpoint and conditions changes it faces in real-world situations. There have been several improvements to the state of the art VPR techniques along with new additions to the list over the last decade. The core problem of VPR is image matching, indeed most of the effort of the research community in recent years has been towards robust image representation techniques. Early approaches were based on hand-crafted image descriptors. SIFT [25] is a local feature descriptor that is used for VPR in [48]. SIFT detects keypoints from an image using Difference-of-Gaussian (DoG) and uses Histogram-of-Oriented-Gradients (HOG) to compute a descriptor of their neighbourhood. SURF is inspired by SIFT but is more efficient and it is used in a variety of VPR approaches [12], [36]. CenSurE [2] is used in FrameSLAM [23]. CenSurE detects keypoints in images using centre surrounded filters across multiple scales at each pixel location. Other important handcrafted techniques are Bag-of-Visual-Words (BOW) [40] and Vector of Locally Aggregated Descriptor (VLAD) [21]. They are used to partition the feature space, such as SIFT descriptors, in a fixed number of visual words in order to obtain a more compact yet informative image representation that allows more efficient image matching. For example, FAP-MAP 2.0 [12] and [16] use BOW and VLAD respectively. Gist [39] is an example of global image descriptor which is used for image matching in [37], [50] and [46]. Gist extracts global features from an image using a set of Gabor filters at different orientations and frequencies. The results are then averaged to obtain an image representation in the form of a vector. Another global descriptor is Histogram-of-Oriented Gradients (HOG) [13], [18]. It calculates the gradient of all image pixels and uses the results to create a histogram, with each bar representing the gradient angles and carrying the summation of gradient magnitudes. McManus *et al.* used HOG for VPR in [31]. SeqSLAM [35] performs the localization by comparing a sequence of images against previously visited places. SeqSLAM does not base the image representation on local or global features but uses intensity patch normalization instead.

In recent years, several VPR approaches based on machine learning have been proposed. The features computed by pre-trained AlexNet can be used for VPR [20]. In particular, the features extracted from the *conv3* layer are most robust to condition variations while those from *pool5* are better for viewpoint changes [57]. The RegionVLAD [22] results in an improvement in image retrieval speed and accuracy due to its low computational CNN-based regional approach combined with VLAD. While CALC [33] is also a convolutional auto-encoder that uses a distorted version of the base image as input and regenerates the HOG descriptor. Next, NetVLAD [3], an advanced version of VLAD that is commonly used for image retrieval, consists of two trainable end-to-end stages. The first is a CNN that extracts the features from an image and the second is a layer that combines them to form an image descriptor by mimicking the behaviour of VLAD. AMOSNet and HybridNet [8] have the same architecture as CaffeNet [24] but they are trained differently. While AMOSNet has been trained from scratch on Specific Places Dataset (SPED) [8], HybridNet uses the weights of the top-5 convolutional layers from CaffeNet, which is trained on ImageNet dataset [43]. Cross-Region-Bow [11] is a cross-convolution technique that collects traits and features from convolutional layers. It further collects the highest 200 energetic regions described using the activations from prior convolutional layers by searching for the prominent sectional approaches from the layers of object-centric VGG-16 [45]. The regional maximum activation of convolutions (R-MAC) [51] operates on the principle that region based description of features can increase matching performance. For CNN based descriptors that proved efficient for image search while the results obtained were improved by employing the geometric re-ranking and query expansion particularly by utilizing the encoded several image regions made by max-pooling.

With the significant addition of VPR techniques, an enquiry that rises in importance is the evaluation of performance differences between these algorithms. Previously, for each technique a different assessment methodology has been proposed, mostly based on Precision-Recall Curves [5], [16], [29], [44]. Ehsan *et al.* [15] present a performance comparison made for evaluating the limitations of image feature detectors utilizing repeatability measure, however, it significantly draws attention to the importance of performance analysis. Among the several VPR techniques mentioned above, only a few have been used for performance comparison before [56] while employing three standard metrics, Matching Performance, Matching Time, Memory Footprint, however, the datasets used in the experimental setup were only a moderate size thus limiting the diversity of the conclusion.

III. Evaluation Framework

The proposed evaluation framework consists of two main phases. The first one determines the upper and lower performance bounds of a VPR method in a given operating environment. This allows determining the performance consistency across an environment. The second phase is designed to compare VPR methods. Comments in [54] suggest that many evaluation approaches tend to emphasize on beating the latest benchmark numbers without considering whether the improvement of vision system over other methods is statistically significant. This consideration can be extended to VPR evaluation where most methods seem to have confined themselves to some particular test conditions to demonstrate their superiority over other competing techniques. Driven by these motivations, the second phase of the evaluation framework uses the McNemars test to determine whether the performance differences are statistically significant or they are due to random artefacts in data. The evaluation framework is based on the new Extended Precision (EP) to measure VPR performance. As detailed below, EP addresses several shortcomings of the existing metrics and it can also be used independently from the proposed framework to evaluate VPR performance.

A. Extended Precision Measure

VPR tasks are characterized by datasets with a prominent skew where positive matches for a query image are rare as compared to negative matches. As Precision-Recall curves

(PR-Curves) are preferable with imbalanced data, they are frequently used to evaluate VPR [14], [41]. Precision (P) and Recall (R) are computed from the outcome of a VPR algorithm: the correct matches are regarded as True Positives (TP) whereas the incorrect matches called as correct are regarded as False Positives (FP). The matches erroneously excluded from the query results are denoted as False Negatives (FN). Precision is the ratio between the correct matches and the total of the predicted positive matches. Recall denotes the proportion of real positives cases that are correctly identified as positive matches. Formally:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \qquad (1)$$

A PR-Curve shows the relation between Precision and Recall and can be obtained by varying an algorithm's parameters [35], the threshold to call a positive match [56] or the number of retrieved images [16]. A PR-Curve can be summarized with several indices. Area Under the PR-Curve (AUC) [14] indicates a VPR performance with a value between 0 and 1. However, AUC does not retain any information regarding the features of the original PR-Curve, including whether Precision reaches or not 1 at any Recall value. R_{P100} [6] is also an important performance indicator; it represents the Recall value at which the Precision drops from 100%, namely it is the highest value of the recall that can be reached without any FP. As a single FP may cause severe failures for many robotic applications [6], [30], R_{P100} is considered a good performance indicator and it is widely employed for VPR evaluation [6], [26]. However, R_{P100} is not capable of determining the lower performance bounds of a VRP method. Indeed, R_{P100} cannot be determined for those PR-Curves that never hit 100% Precision. To circumvent this problem we introduce Extended Precision:

$$EP = \frac{P_{R0} + R_{P100}}{2} \qquad (2)$$

where P_{R0} denotes the precision at the minimum Recall value and the factor '2' in the denominator is to have $EP \in [0, 1]$. If $P_{R0} < 1$, R_{P100} is set to 0 and EP depends only on the Precision at minimum Recall while for $P_{R0} = 1$, EP is greater than 0.5 and works similarly to R_{P100}. An example is given in Fig. 1. VPR1 has $P_{R0} = 1$ and $R_{P100} = 0.6$ then, the corresponding EP is 0.8. The Precision of VPR2 is constantly below 1 thus R_{P100} is undefined end set to 0. The resulting EP for VPR2 is 0.4. Accordingly with EP definition, VPR1 outperforms VPR2.

EP combines P_{R0} and R_{P100} into a single scalar value which provides more comprehensive insights into VPR performance than using them individually. P_{R0} is determined by the number of FPs before the first TP and can only measure the precision of a method, without describing how the performance is affected by including more query results. R_{P100} indicates the occurrence of the first FP but its applicability requires $P_{R0} = 1$. Therefore, it cannot be computed for any PR-Curve and it cannot measure VPR lower performance bounds.

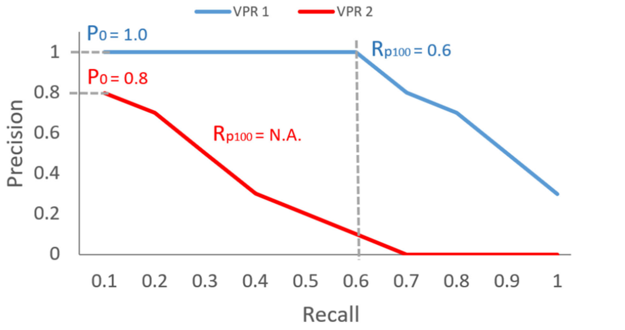

Fig. 1. An example of how our proposed EP is computed for two hypothetical VRP systems. At glance, the two curves suggest that VRP1 is better than VPR2. Indeed, the corresponding EP values are 0.8 and 0.4 respectively.

B. Identification of the Upper and Lower Performance Bounds

In the first phase of the proposed framework the upper and lower performance bounds of VPR techniques are identified. The evaluation process uses two sets of images. A reference dataset (I_{REF}) which represents the previously visited places and a query dataset (I_e) that shows the same locations as the reference dataset but under different viewing conditions (e.g., from a different viewpoint). This phase consists of the following steps which are repeated for every VPR method to assess.

i) Let v be a VPR technique and q a query image. The images in I_{REF} are ranked by their similarity with q using v. Then, a PR-Curve for q is computed by varying the number of retrieved images from 1 to the last image that corresponds to $R = 1$. For each step, a confusion matrix is computed using the ground truth, the corresponding P and R values form a point of the PR-Curve. This process is repeated for all $q \in I_e$ to produce a set of curves for method v.

ii) For each PR-Curve from the step i), the pair (P_{R0}, R_{P100}) is computed. Then equation (2) is used to compute the set of EP values for v:

$$E_v = \{EP_1, EP_2, \ldots\ldots EP_n\} \qquad (3)$$

iii) The upper and lower performance bound for v on the dataset I_e corresponds with the highest and lowest EP values in E_v respectively:

$$EP_{Max} = \max(E_v), \quad EP_{\min} = \min(E_v) \qquad (4)$$

iv) The proposed approach considers the precision crucial for VPR as FPs might have a severe impact on robotic applications [6]. To this end, the ratio of query images with $EP > 0.5$ is a relevant performance indicator:

$$S_{P100} = \frac{|\{i \in I_e | EP > 0.5\}|}{n} \qquad (5)$$

where n is the number of images in I_e.
It is worth mentioning that when VPR is cast as an image retrieval task, $EP > 0.5$ indicates that the first retrieved image ranked by similarity is a correct match. Thus, S_{P100} represents the share of successful single matches in I_e and

it is particularly useful to assess VPR-based systems that use a single image to perform localization.

C. Identification of Statistically Significant Performance Differences

This part of the proposed evaluation framework is concerned with determining whether the performance differences between VPR techniques are statistically significant or are due to random artefacts in data. Following the approach described in [15], the proposed evaluation method interprets the process of testing VPR against a sequence of query images as a series of success/failure trails on the same dataset. Under this assumption, the resulting distribution follows a binomial model and the comparison between two algorithms (v and w) can be addressed with McNemar's test [17], [32].

$$\chi^2 = \frac{(|N_{sf} - N_{fs}| - 1)^2}{N_{sf} + N_{fs}} \tag{6}$$

where the '-1' in the numerator is a continuity correction; N_{sf} denotes the number of trials where the algorithm v succeeded and w failed; N_{fs} denotes the number of trials where v failed and w succeeded. The proposed framework uses Z score which is obtained as the square root of equation (6):

$$Z = \frac{|N_{sf} - N_{fs}| - 1}{\sqrt{N_{sf} + N_{fs}}} \tag{7}$$

When $N_{sf} + N_{fs} \geq 30$, the test is reliable and χ^2 has a chi-squared distribution with one degree of freedom. As with the Chi-Square test, the cut-off point for 95% significance level is 3.84 which corresponds to 1.96 for Z. Therefore, if the Z value is larger than 1.96, one can say the results are a consequence of artefacts in data by chance only one in twenty ($p = 0.05$). McNemar's test cannot be used to compare more than two VPR methods at the same time thus, a series of pairwise comparisons are made. However, executing multiple statistical tests requires a correction to the significance level of each single tests. Bonferroni correction is a well-known solution to this problem: let α the significance level for the whole family of N tests then each test needs to be performed with a significance level of α/N.

To perform the McNemar's test, it is required to determine when an algorithm fails or succeeds. In the proposed framework, success occurs when EP is greater than a threshold t otherwise, a fail is accounted. EP is characterized by two intervals: 0 to 0.5 where the value is determined by P_{R0}, and 0.5 to 1 where EP mimics the behaviour of R_{P100}. This feature of EP allows comparing VPR methods from different perspectives by using multiple thresholds. If McNemar's test is performed with a threshold ≤ 0.5, the VPR pair are compared on the basis of P_{R0}, which is determined by the number of FPs before the first occurrence of a TP. Conversely, using thresholds greater than 0.5 successes and failures are determined by R_{P100}, namely by the length of the TP sequence before the first FP occurrence in the retrieved images.

Let T be the set of thresholds used to execute the McNemar's test variant:

$$T = \{t_1, t_2, \ldots, t_p\}, \quad \forall i = 1, 2, \ldots, p \tag{8}$$

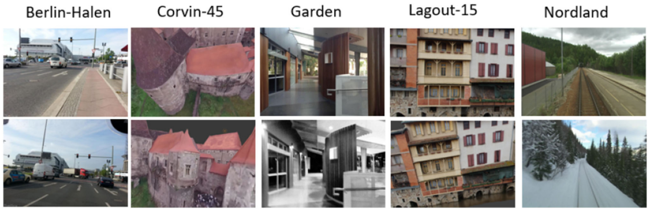

Fig. 2. A sample of the datasets used for the experiment. Reference images at the top and query images at the bottom.

For a pair of VPR methods to compare, a set of Z scores is computed using the equation (7), one for each value $t_i \in T$.

$$Z_{vw} = \{z_1, z_2, \ldots, z_t\} \tag{9}$$

where v and w denote the tested VPR techniques and z_i is the value of Z obtained with the i^{th} threshold value in T. Although there is not any specific selection criterion for T, a good practice is to select the threshold values in order to capture the entire spectrum of variations of the performance metric [15]. As detailed in Section IV, a good setup for EP is with nine evenly spaced thresholds between 0.1 and 0.9.

IV. Results

The proposed evaluation framework is employed to compare several state-of-the-art VPR methods: AMOSNet, HybridNet [8], R-MAC [51], NetVLAD [3] and Cross-Region-Bow [11]. To test AMOSNet and HybridNet, we used the models trained with SPED dataset [8] by their authors [9]. The implementation of R-MAC used for the experiments has been obtained from [52]. For a fair comparison, the geometric verification module has been deactivated for the tests. The MATLAB source of NetVLAD is available from [4] along with several sets of weights. The results presented in this section are obtained using the VGG-16 model trained with Pittsburgh 250 K dataset [53] and using a dictionary with 64 words. Cross-Region-Bow is also available as a MATALB implementation [10]. For the experiments, the VGG-16 model pre-trained on ImageNet dataset [43] has been utilized with a BoW dictionary of 10 K words.

In order to obtain comprehensive results, VPR methods have been assessed under different image variations using the five datasets summarized in Table I and shown in Fig. 2. Berlin Halensee Strasse [11] includes two traverses of an urban environment. This dataset exhibits moderate to strong viewpoint variations and changes in appearance due to dynamic elements such as cars and pedestrians. The ground truth is obtained using GPS coordinates to build place-level correspondence using a maximum distance of 25 meters as a criterion. For the experiments, the image set berlin-halenseestrasse-1 has been used as a reference and berlin-halenseestrasse-2 as a query dataset. Lagout and Corvin [29] are synthetic datasets consisting of several flybys around buildings. Lagout traverses at 0° and 15° are used as reference and query datasets respectively to test VPR techniques under moderate viewpoint changes. Similarly, the Corvin's loops captured at ground level and at 45° are used to assess VPR methods under very strong viewpoint changes. The

TABLE I
DATASET VARIATIONS AND GROUND TRUTH TOLERANCE

Dataset	Variation		Ground Truth Tolerance
	Viewpoint	Condition	
Berlin Halensee Strasse	Strong	Dymamic Objects	25m
Corvin 45	Very Strong	None	GT available
Garden Point	Mild	Day - Night	±2 frames
Lagout 15	Moderate	None	GT available
Nordland	None	Seasonal	±5 frames

ground truth data for Lagout and Corvin are made available by their authors [1]. Gardens Point Dataset [8] consists of three traverses of the Queensland University of Technology (QUT). Two occurred during the daylight by walking on two opposite sides of the walking path (laterally changed viewpoint) and the third during the night on the right side. The results are presented for illumination changes, thus the right-day and right-night traverses are used as reference and query datasets respectively. The traverse footages are synchronized thus the ground truth is obtained by frame correspondences. For the test, a reasonable tolerance is to consider a match correct if the query and the retrieved images are within 5 frames from each other [26], that is a retrieved reference image must fall between $n-2$ and $n+2$ where n is the query image index. Nordland Dataset [47] is built from footage for every season along a railroad in Norway. It shows extreme seasonal changes, especially between summer and winter journeys, which are used as reference and query datasets to obtain the results presented in this letter. Similarly to Garden Points, the footages are synchronized but the speed of the train is considerably faster than a human walk. Thus, the ground truth is built considering a tolerance of 11 frames as indicated in [26]. That is a reference image must fall between $n-5$ and $n+5$.

A. Upper and Lower Performance Bounds Discussion

The results obtained with the use of the first phase of the proposed framework are summarized in Fig. 3. Green bars represent the upper performance bounds of VPR methods and correspond to EP_{Max}. Similarly, the lower performance bounds $EP_{\min}$ are represented with red bars. The values of S_{P100} are indicated in yellow and are read on the right-side y-axis.

In terms of EP_{Max} the considered VPR methods exhibit comparable performance on Berlin, Garden and Lagout with EP values equal or close to 1. Thus, all the VPR techiniques reach a good performance peak with dynamic objects, illumination and moderate to strong viewpoint changes. The prominent viewpoint variations of Corvin pose a hard challenge and none of the tested methods can reach $EP = 1$. In every Corvin's location, the assessed methods cannot recover all the true matches for a query image without including one or more FPs in the result set. $EP_{Max} < 1$ indicates that there are not easily recognizable places that a robot system can use to localize itself reliably. Similarly, Nordland is also a difficult environment as we used the summer and winter traverses that exhibit prominent variations in appearance. Indeed, except for NetVLAD, EP never hit 1. This is due to datasets used to train the considered VPR which are not meant to cope with extreme seasonal variations.

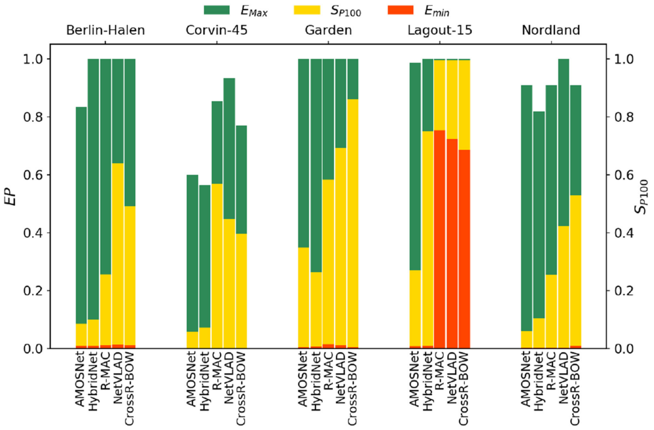

Fig. 3. Upper and lower performance bounds for the assessed VPR methods. The green and red bars represent the maximum and minimum EP scored on each dataset (top x-axis). The yellow bars indicate S_{P100} and the related y-axis is on the right side of the figure.

S_{P100} indicates the place share where a VPR technique successfully identifies a true match as the most similar image to the query. From the perspective of S_{P100}, the differences between VPR methods are more significant. NetVLAD is the best approach in the urban environment of Berlin-Halensee. This is not surprising as the model has been trained using images captured from urban scenes. Cross-Region-Bow appears to be the most reliable VPR method for illumination and seasonal changes. It scores a S_{P100} of 0.88 and 0.53 on Garden Point and Nordland respectively. Cross-Region-Bow uses a pre-trained network on ImageNet which is not prominent in any specific image transformation. Thus, its performance should be accounted to the approach used to combine features into a robust image representation. Corvin is confirmed to be the most difficult dataset also from the perspective of S_{P100}. The only technique that can hit at least $S_{P100} = 0.5$ is R-MAC, which can be considered the most reliable VPR method on Corvin.

The lowest performance bound is close to zero in most of the tested scenarios. This means that in some places the localization by means of visual features might be very difficult because of the frequent occurrence of FPs in the retrieved image set. The only exceptions are R-MAC, NetVLAD and Cross-Region-Bow whose lower bounds are constantly above 0.5 on Lagout. As $EP_{\min} \geq 0.5$ requires $P_{R0} = 1$, the most similar reference image to the query which is retrieved by these three methods is a TP in every Lagout's place.

B. Statistical Performance Comparison

This section presents a statistical performance comparison between the VPR methods evaluated in the previous phase. The results are obtained by utilizing the second phase of the proposed framework and presented in Fig. 4. The threshold set T used for the experiments includes 9 values ($p = 9$) equally spaced

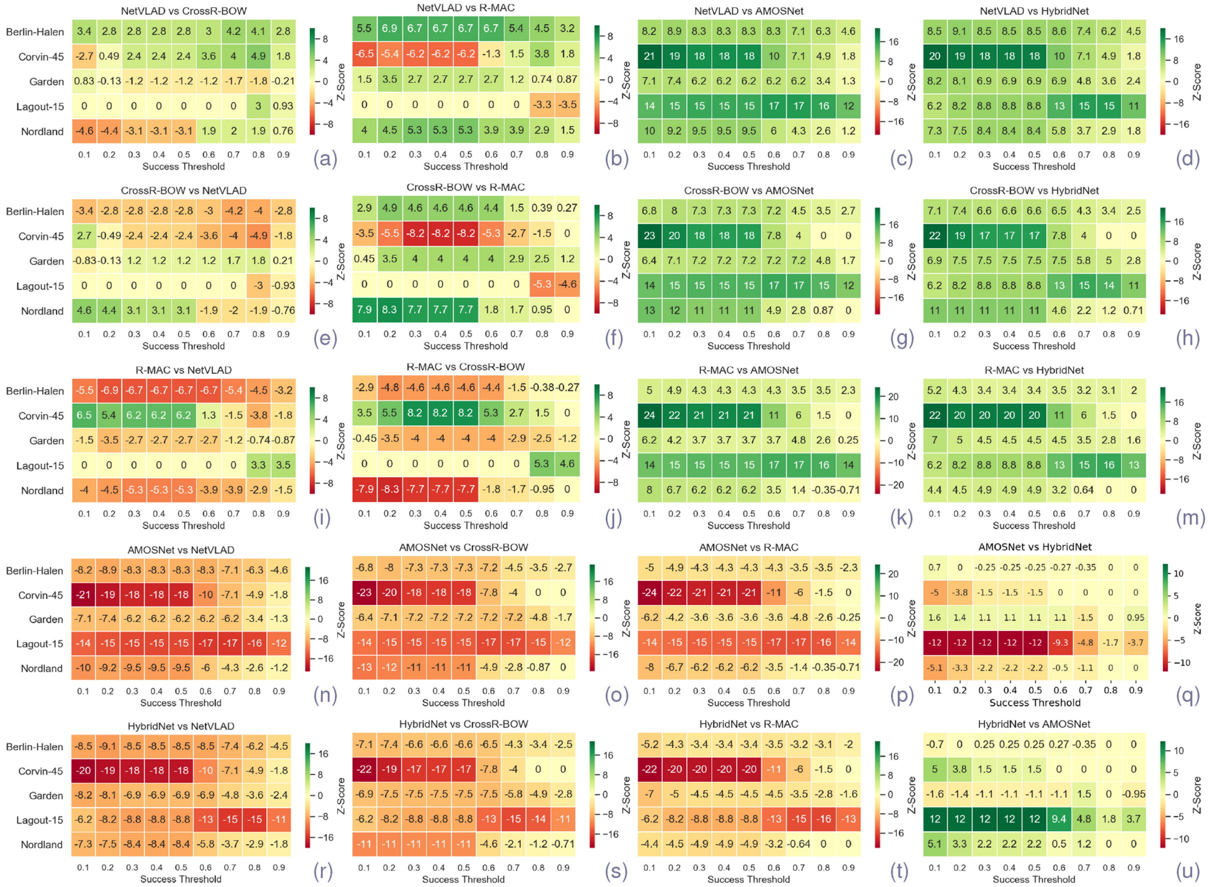

Fig. 4. Pair-wise comparison for the VPR methods tested. A sign convention is used to present the results, a positive value of Z indicates that the first method of the pair outperforms the second one, whereas a negative Z score has the opposite meaning.

between 0.1 to 0.9. As detailed in Section III-C, at low threshold values a successful trial is determined by the component P_{R0}. Conversely, for thresholds greater than 0.5, is the component R_{P100} of EP to determine successes and fails. This setup allows the comparison of VPR methods from different perspectives by exploring the complete range of variation of EP. In Fig. 4 a colour code is used to represent the Z values for all combinations of threshold and dataset. Although Z is always positive, a sign convention is used to indicate which VPR method obtains better performance. A positive Z score means that the first technique of the pair is better than the second, namely $N_{sf} > N_{fs}$. A negative value of Z indicates that is the second VPR method of the pair to outperform the first one ($N_{sf} < N_{fs}$). It is worth noting that $|Z|$ increases with the difference $|N_{sf} - N_{fs}|$ thus, Z can be interpreted as a performance gap between two VPR approaches.

NetVLAD and Cross-Region-Bow outperform the other approaches on Berlin-Helensee, Graden Point and Nordland datasets as confirmed by their large positive Z values (Fig. 4(b) to Fig. 4(h)). They have comparable performance on Corvin and Lagout while NetVLAD is better than Cross-Region-Bow on Berlin-Halenstrasse and worse on Nordland at every threshold. (Fig. 4(a) and Fig. 4(e)). HybridNet outperforms or achieves comparable performance as AMOSNet in most of the test scenarios (Fig. 4(u)). Our results is coherent with the performance analysis by their authors [8]. Z-score exhibits wide variations on Corvin for every VRP technique. In particular, R-MAC presents large positive Z values for thresholds between 0.1 and 0.5 against every other VPR technique (third row in Fig. 4) At larger thresholds, Z decreases and becomes negative against NetVLAD starting from 0.7 (Fig. 4(i)) Thus, R-MAC outperforms the other approaches when the evaluation is carried out by observing low EP values which are mostly influenced by P_{R0}. As the threshold increases, the number of successes is determined by the contribution of R_{P100}. In such evaluation conditions, R-MAC is outperformed by NetVLAD which demonstrates to be capable of retrieving longer sequences of TPs on Corvin. McNemer's test outcome confirms and supports the bounds analysis presented in the previous section. As it is shown in 3, R-MAC has the best S_{p100} while Cross-Region-Bow and NetVLAD reach higher EP_{Max}.

C. AUC as an Alternative to EP

AUC can be used as an alternative to EP to measure VPR performance. However, we consider AUC less appropriate than Extended Precision for use in the proposed evaluation framework. The most important reason is that AUC does not penalize top-ranked FPs in the query results. Indeed, AUC might be significantly incremented by long sequences of TPs regardless of their position in the retrieved image ranking. As opposed to

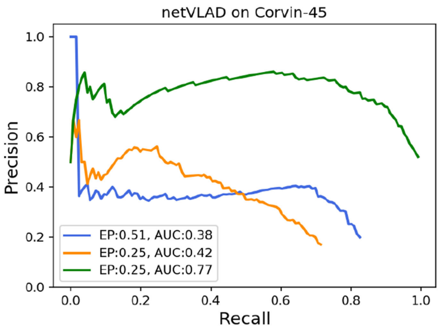

Fig. 5. A comparison between three PR-Curve for netVLAD on Corvin with their respective EP and AUC values.

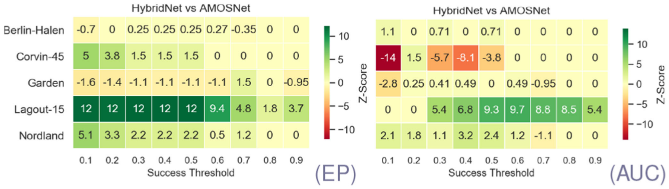

Fig. 6. McNemer's test using EP (left) and AUC (right) to compare HybridNet and AMOSNet.

this, P_{R0} component of Extended Precision penalizes top-ranked false positives by forcing $EP \leq 0.5$. In other words, large AUC does not guarantee the first images retrieved by a VPR technique are correct matches. For example, the blue curve in Fig. 5 has AUC $= 0.38$ and $EP = 0.51$. The green curve has a larger AUC (0.77) and a smaller EP (0.25). As described in Section III-B.iv, $P_{R0} = 1$ is an important evaluation criterion for the proposed evaluation framework, thus the blue curve is considered better than the green one regardless of the smaller AUC.

Finally, AUC is more difficult to interpret than P_{R0}. Except for 0 and 1, the value of AUC is not related to any specific condition or PR-Curve feature. For this reason, McNemar's test based on AUC is harder to understand. Fig. 6 shows the test results for a pair of VPR methods using both EP and AUC. The large negative score of HybridNet against AMOSNet at 0.5 on Corvin means that HybridNet's AUC cannot reach the threshold as often as AMOSNet. However, a clear interpretation of this outcome is hard to give as AUC ≥ 0.5 does not have any specific meaning related to VPR performance. Conversely, the positive Z value at 0.5 for the EP-based test means that the top-1 retrieved image by HybridNet on Corvin is more often a correct match than for AMOSNet.

V. Conclusion

In this letter, a new framework to evaluate VPR performance is proposed. It consists of two phases: one is designed to assess the consistency of a VPR method performance across an environment and the second uses a variant of McNemar's test to identify the statistically significant performance differences between VPR methods. The proposed framework is based on the newly introduced Extended Precision measure for VPR performance. EP summarizes a PR-Curve by combining two of its most relevant features, P_{R0} and R_{P100}, into an easy to read measure for VPR performance. EP addresses several shortcomings of AUC which would produce less significant and hard to understand results if used with the proposed evaluation method. The proposed framework is then used to assess and compare several state-of-the-art VPR techniques using different datasets including one or more appearance variations such as illumination and viewpoint changes. NetVLAD has shown solid end reliable performance in most of the test scenarios and in urban environments in particular. Cross-Region-Bow has exhibited good performance too, especially with illumination and seasonal variations. AMOSNet and HybridNet achieved the worst performance among the considered methods, especially in dealing with strong viewpoint variations where, to the contrary, R-MAC resulted to be the most reliable VPR approach.

References

[1] V4RL Wide-baseline Place Recognition Dataset, 2019. [Online]. Available: https://github.com/VIS4ROB-lab/place_recognition_dataset_ral2019, Accessed: Apr. 4, 2019.

[2] M. Agrawal, K. Konolige, and M. R. Blas, "Censure: Center surround extremas for realtime feature detection and matching," in *Proc. Euro. Conf. Comput. Vis.*, 2008, pp. 102–115.

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[4] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "netvlad implementation," 2016. [Online]. Available: https://github.com/Relja/netvlad, Accessed: Sep. 4, 2019.

[5] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Towards life-long visual localization using an efficient matching of binary sequences from images," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 6328–6335.

[6] D. Bai, C. Wang, B. Zhang, X. Yi, and X. Yang, "Sequence searching with CNN features for robust and fast visual place recognition," *Comput. Graph.*, vol. 70, pp. 270–280, 2018.

[7] M. Bürki *et al.*, "VIZARD: Reliable visual localization for autonomous vehicles in urban outdoor environments," *IEEE Int. Veh. Symp. (IV)}*, 2019, pp. 1124–1130, *arXiv:1902.04343*.

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

[9] Z. Chen *et al.*, "Amosnet and hybridnet implementation," 2017. [Online]. Available: https://github.com/scutzetao/DLfeature_PlaceRecog_icra2017, Accessed: Sep. 4, 2019.

[10] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Cross-region-bow implementation," 2016. [Online]. Available: https://github.com/scutzetao/IROS2017_OnlyLookOnce, Accessed: Sep. 4, 2019.

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

[12] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *Int. J. Robot. Res.*, vol. 30, no. 9, pp. 1100–1123, 2011.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[14] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd ACM Int. Conf. Mach. Learn.*, 2006, pp. 233–240.

[15] S. Ehsan *et al.*, "A generic framework for assessing the performance bounds of image feature detectors," *Remote Sens.*, vol. 8, no. 11, p. 928, 2016. [Online]. Available: https://www.mdpi.com/2072-4292/8/11/928

[16] B. Ferrarini, M. Waheed, S. Waheed, S. Ehsan, M. Milford, and K. D. McDonald-Maier, "Visual place recognition for aerial robotics: Exploring accuracy-computation trade-off for local image descriptors," in *Proc. NASA/ESA Conf. Adaptive Hardware Syst.*, 2019, pp. 103–108.

[17] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*. Hoboken, NJ, USA: Wiley, 2013.

[18] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *Proc. Int. Workshop Autom. Face Gesture Recognit.*, 1995, vol. 12, pp. 296–301.

[19] S. Garg, N. Suenderhauf, and M. Milford, "Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics," in *Proc. Robot.: Sci. Syst.*, Jun. 2018, doi: 10.15607/RSS.2018.XIV.022.

[20] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sens.*, vol. 9, no. 8, p. 848, 2017. [Online]. Available: https://www.mdpi.com/2072-4292/9/8/848

[21] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3304–3311.

[22] A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight CNNs for severe viewpoint and appearance changes," 2019, pp. 1–9, doi: 10.1109/TRO.2019.2956352.

[23] K. Konolige and M. Agrawal, "FrameSLAM: From bundle adjustment to real-time visual mapping," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1066–1077, Oct. 2008.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

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

[27] S. Lowry *et al.*, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[28] F. Maffra, Z. Chen, and M. Chli, "Viewpoint-tolerant place recognition combining 2D and 3D information for UAV navigation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 2542–2549.

[29] F. Maffra, L. Teixeira, Z. Chen, and M. Chli, "Real-time wide-baseline place recognition using depth completion," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1525–1532, Apr. 2019.

[30] M. Magnusson, H. Andreasson, A. Nuchter, and A. J. Lilienthal, "Appearance-based loop detection from 3D laser data using the normal distributions transform," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 23–28.

[31] C. McManus, B. Upcroft, and P. Newmann, "Scene signatures: Localised and point-less features for localisation," in *Proc. Robot.: Sci. Syst.*, Jul. 2014, doi: 10.15607/RSS.2014.X.023.

[32] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[33] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," in *Proc. Robot., Sci. Syst.*, Pittsburgh, PA, USA, Jun. 2018, doi: 10.15607/RSS.2018.XIV.032.

[34] M. Milford *et al.*, "Sequence searching with deep-learnt depth for condition-and viewpoint-invariant route-based place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 18–25.

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

[36] A. C. Murillo, J. J. Guerrero, and C. Sagues, "Surf features for efficient robot localization with omnidirectional images," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2007, pp. 3901–3907.

[37] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops*, 2009, pp. 2196–2203.

[38] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2564–2570.

[39] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Prog. Brain Res.*, vol. 155, pp. 23–36, 2006.

[40] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[41] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Machine Learn. Tech.*, vol. 2, no. 1, pp. 37–63, 2011.

[42] A. Ranganathan, S. Matsumoto, and D. Ilstrup, "Towards illumination invariance for visual localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 3791–3798.

[43] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[44] N. V. Shirahatti and K. Barnard, "Evaluating image retrieval," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 955–961.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[46] G. Singh and J. Kosecka, "Visual loop closing using gist descriptors in manhattan world," in *Proc. IEEE Int. Conf. Robot. Autom. Omnidirectional Vis. Workshop*, 2010, pp. 44–48.

[47] S. Skrede, "Nordlandsbanen: Minute by minute, season by season," 2013. [Online]. Available: https://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-sea son/

[48] E. Stumm, C. Mei, and S. Lacroix, "Probabilistic place recognition with covisibility maps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 4158–4163.

[49] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons," in *Proc. IEEE Int. Conf. Robot. Autom. Workshop Long-Term Autonomy*, 2013, p. 2013.

[50] N. Sünderhauf and P. Protzel, "Brief-gist-closing the loop by simple means," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2011, pp. 1234–1241.

[51] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," May 2016, pp. 1–12.

[52] G. Tolias, R. Sicre, and H. Jégou, "R-MAC implementation," 2016. [Online]. Available: https://github.com/gtolias/rmac, Accessed: Sep. 4, 2019.

[53] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 883–890.

[54] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1521–1528.

[55] C.-C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, "Simultaneous localization, mapping and moving object tracking," *Int. J. Robot. Res.*, vol. 26, no. 9, pp. 889–916, 2007.

[56] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions," 2019, *arXiv:1903.09107*.

[57] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.