

What makes visual place recognition easy or hard?

Stefan Schubert and Peer Neubert

TU Chemnitz, Germany

{stefan.schubert, peer.neubert}@etit.tu-chemnitz.de

Abstract—Visual place recognition is a fundamental capability for the localization of mobile robots. It places image retrieval in the practical context of physical agents operating in a physical world. It is an active field of research and many different approaches have been proposed and evaluated in many different experiments. In the following, we argue that due to variations of this practical context and individual design decisions, place recognition experiments are barely comparable across different papers and that there is a variety of properties that can change from one experiment to another. We provide an extensive list of such properties and give examples how they can be used to setup a place recognition experiment easier or harder. This might be interesting for different involved parties: (1) people who just want to select a place recognition approach that is suitable for the properties of their particular task at hand, (2) researchers that look for open research questions and are interested in particularly difficult instances, (3) authors that want to create reproducible papers on this topic, and (4) also reviewers that have the task to identify potential problems in papers under review.

I. INTRODUCTION

The task of visual place recognition (VPR) is to find associations between one or multiple query images with a database of images of known places. It is an important means for mobile robot localization, in particular for loop-closure detection during SLAM or for candidate selection for pose estimation. Due to its importance for long-term operation, a particularly active research field is recognizing places despite severe changes of the environmental conditions induced by day-night cycles, changing weather, or seasons. At its core, VPR is an image retrieval problem and many developments and results from the computer vision community (e.g. image descriptors) significantly contributed to the state of the art. However, the application in the context of mobile robotics frames the image retrieval task and typically provides additional structure and information (e.g. spatio-temporal consistency). This framing had two important effects for the research in this area: First, adapting to the different framing led to the emergence of own datasets and evaluation best-practices in the robotics community which are different to the computer vision community. In particular, the standard datasets and benchmarks from computer vision are typically not applied. Second, dependent on the additional structure and information that is available and exploited, significantly different types of place recognition problem setups emerged. This leads to a huge variety and volatility of experimental setups and evaluation procedures among different papers. Which then leads to barely comparable results between these papers and a lack of common understanding of the state of

the art. The negative consequences include a fragmentation of the community and severe impediments for monitoring the advances in the field and for identification and structured addressing of open research questions.

In this article, we provide an overview of the different existing place recognition problem setups. Starting point is a presentation of a basic place recognition pipeline in Sec. II. This is used to demonstrate the influence of design decisions on the comparability of results across groups and papers in Sec. III. We will then provide an extensive list of properties of place recognition problems that can have large influence on the difficulty of the task in Sec. IV. Finally, we will identify particularly challenging problem setups and property combinations that are interesting for future research in Sec. V.

II. PRELIMINARIES: A BASIC VISUAL PLACE RECOGNITION (EVALUATION) PIPELINE

Visual place recognition is an important subproblem of mobile robot localization and subject of intense research, please refer to [1], [2] for surveys. It is required for tasks like loop closure detection in SLAM or candidate selection for global localization. Fig. 1 shows a basic place recognition pipeline. Input are database and query sets, the output is a set of (candidate) matchings between the two sets. The basic computing steps include (1) computation of descriptors, (2) pairwise comparison of image descriptors to generate a (potentially sparse) similarity matrix and (3) potentially the deduction of a set of discrete matchings (or matching candidates) between the two image sets. The basic pipeline can be extended with additional pre- and postprocessing steps, e.g. for descriptor standardization [3] or sequence processing [4].

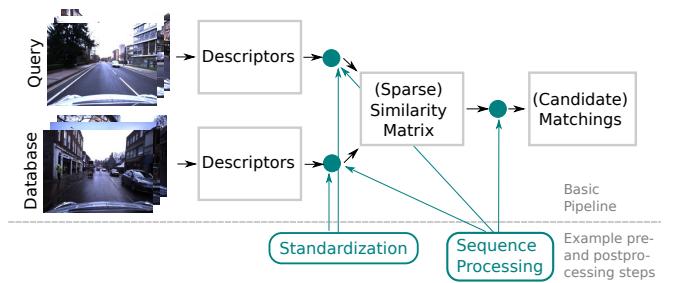


Fig. 1. Even this very basic place recognition pipeline with descriptor and similarity matrix computation has many degrees of freedom that need to be specified. This becomes even more complex if it is extended with pre- and postprocessing steps.

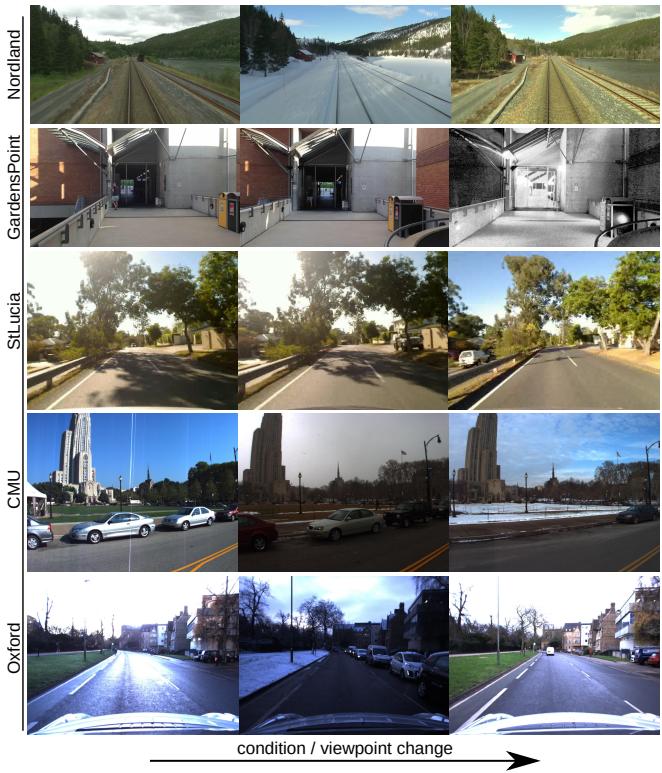


Fig. 2. Example images of same places under different conditions for five standard datasets.

However, not all place recognition approaches follow this structure. For example, end-to-end deep-learned methods (e.g. PoseNet-like approaches [5]) encode information about the database in their neural network weights and infer the image pose from a single query image without any explicit descriptor or image comparison.

Nevertheless, we want to use this basic pipeline to discuss the comparability of place recognition experiments and to identify their distinguishing properties. More details on the individual parts of this basic pipeline can be found in Appendix A. If Fig. 2 makes sense to you, there is probably no need to read this appendix, otherwise, it might be helpful to have a look.

Foundation for performance evaluation and comparison of place recognition approaches is ground-truth knowledge about matchings between database and query images. Similar to other robotics applications, the generation of datasets with ground-truth information is a laborious task. However, in case of place recognition, we face an additional (and potentially more severe) problem: Place recognition is an ill-defined problem and there is no inherently correct or wrong answer to the question whether two images show the same place or not. For example, an image of the Eiffel tower and another image of Notre-Dame cathedral can be considered to show *different* places within the city of Paris, or to show the *same* place (“Paris”) within France. To account for this problem, the creation of ground-truth information for place recognition (or more precisely, creation of the target output of the place recognition algorithm), typically adds additional

constraints to the problem, e.g. on the maximum allowed spatial distance of camera poses or difference between image indexes.

Table I shows a collection of datasets that have been used for the evaluation of place recognition algorithms, Fig. 2 shows some example images. They provide a large variety of properties of database and query sets including the type of environment (e.g., indoor/outdoor, urban/rural), type of condition change between database and query (e.g., dynamics, weather, seasonal), constant conditions *within* database and query or not, viewpoint changes, minimum/maximum spatial distance between consecutive images (e.g., with/without stops), loops within database or query, and so on. Many datasets are publicly available and were repeatedly used in different works. However, due to the complexity of some of the datasets and potential ambiguities in the ground-truth place matchings, the comparison of results on a (at first glance) same dataset from different papers is unreliable. This will be discussed in more detail in the following sections.

Due to the large variety of potential environments and environmental conditions, it is good practice to evaluate algorithms on a combination of available datasets. However, this choice opens opportunities to cherry-pick datasets where the preferred algorithm performs best. There have been a few attempts to provide a systematic collection of datasets and to establish a standard benchmark. They are shortly reviewed in the Appendix B. However, so far there is no established standard for the evaluation of place recognition algorithms. This leads to the fundamental problem which is discussed in the next section.

III. PLACE RECOGNITION RESULTS ARE BARELY COMPARABLE ACROSS PAPERS

How can I select the best place recognition approach for my task at hand? Can I just look at the most recent papers and choose the algorithm with the best performance measure? Unfortunately, it’s not that simple. In this section, we show some examples, where differences in the experimental setups between different papers can have a huge influence on the results and make comparisons across papers largely impossible. The most obvious is the previously mentioned choice of datasets, but there are others, more subtle influences. The following Sec. IV will then provide a more extensive list of properties that allow to distinguish different place recognition problems and that have to be taken care of to avoid such differences.

A. Choice of the dataset

Table I shows a selection of datasets that have been used for mobile robot place recognition experiments. Of course, the performance of an algorithm can significantly vary between datasets. However, and very importantly, also the relative performance of two algorithms can switch between different datasets. Often, different algorithms will perform best on different datasets.

For example, table II shows results presented in [3] of the AlexNet-based image descriptor from [31] and NetVLAD

TABLE I
A SELECTION OF DATASETS THAT HAVE BEEN USED FOR PLACE RECOGNITION IN MOBILE ROBOTICS.

Dataset	Ref	Appearance Change	Setup	#Seq	GT
Alderley Day/Night	[6]	day, night/rain	car	2	ID
CMU	[7]	time, 4-seasons	car	16	raw GPS
Freiburg Across Seasons	[8]	summer/sunny, winter/low-sun, long-term change	car	3	ID
Freiburg-Bonn	[9]	summer-winter, morning-evening	car	2x2	ID
Gardenspoint Walking	[10]	day-night, viewpoint	hand-held	3	ID
NCLT	[11]	sunny/cloudy, time, seasons	robot	27	poses
Nordland	[12]	seasons, no viewpoint	train	4	ID
Oxford RobotCar	[13]	time, weather, seasons	car	>100	raw GPS
SFU Mountain Dataset	[14]	time, day-night, weather, seasons	robot	7	ID
South Bank Bicycle	[10]	day-night, viewpoint	bicycle	2	ID
StLucia	[15]	time of day	car	10	raw GPS
Symphony Lake	[16]	illumination, seasons	boat	121	poses
V4RL Urban Place	[17]	viewpoint	hand-held	3	ID
All day	[18]	day-night, time, dusk, dawn	car	6	raw GPS
City Center	[19]	dynamics, shadows	robot	1	ID
DIRD	[20]	time	car	2	not available
Ford Campus	[21]	sunny, overcast	car	2x1	raw GPS
Kelvin Grove Footpath	[22]	day-night	hand-held	2	ID
Kitti Odometry	[23]	dynamics	car	11	poses
Malaga Urban Dataset	[24]	dynamics	car	1	raw GPS
Mapillary Berlin	[25]	dynamics, viewpoint	bicycle, bus	2	raw GPS
New College	[19]	dynamics	robot	1	ID
Örebro Seasons	[26]	seasons, weather	robot	7	poses
PitOrlManh	[27]	dynamics, 3 cities	street view	3x1	poses
RAWSEEDS	[28]	time	robot	12+5	poses
StLucia Vision Dataset	[29]	none (no loop)	car	1	raw GPS
Surfers Paradise	[30]	day/rain-night	car	2	ID

[32] descriptors using a simple pairwise image comparison on the Oxford [13] and Nordland [12] datasets. The performance of the raw and the standardized NetVLAD descriptors is significantly better for each comparison from Oxford. In contrast, for each Nordland comparison the AlexNet-based descriptor is better. With knowledge about the details of the algorithms and datasets, we can speculate about the reasons: The constant viewpoint in Nordland presumably helps the flattened-feature-map-style AlexNet descriptor that includes a lot of spatial information; and the domain of the training data for NetVLAD (Pittsburgh 250k) is significantly more similar to Oxford than to Nordland. This example demonstrates that the choice of evaluation datasets can considerably influence the result. This allows to (unintentionally) cherry-pick datasets in favour of a particular algorithm.

B. Details of the usage of the particular dataset

Even if the same image source is used, the exact way how a dataset is used can have a huge influence. A practically relevant example is to use only a part of a specific dataset. For instance, Fig. 3 shows precision-recall curves of the *same* algorithm (AlexNet [31] pairwise as before) using the *same* basic dataset (Nordland spring-winter [12]) but *different fractions* of the trajectory. The thick red curve is obtained when equally sampling images over the whole trajectory. For the others the same number of places is used, but each time sampled only from a smaller part of the trajectory - resulting in a wide variety of curves. Other examples

TABLE II
EVALUATION OF THE AVERAGE PRECISION OF ALEXNET AND NETVLAD DESCRIPTORS ON NORDLAND AND OXFORD DATASETS, EITHER USING RAW DESCRIPTORS OR WITH STANDARDIZATION [3]. THE COLORED ARROWS INDICATE LARGE ($\geq 25\%$) OR MEDIUM ($\geq 10\%$) DEVIATION FROM “RAW”. SEE [3] FOR DETAILS.

Dataset	Reference	Query	NetVLAD		AlexNet	
			Raw	STD	Raw	STD
Nordland	fall	spring	0.39	0.61 ↑	0.82	0.92 →
	fall	winter	0.06	0.26 ↑	0.63	0.83 ↑
	spring	winter	0.11	0.37 ↑	0.59	0.86 ↑
	summer	spring	0.32	0.58 ↑	0.77	0.90 →
	summer	fall	0.63	0.84 ↑	0.94	0.97 →
	141209	141216	0.87	0.92 →	0.49	0.65 ↑
Oxford	141209	150203	0.93	0.96 →	0.63	0.85 ↑
	141209	150519	0.83	0.91 →	0.25	0.80 ↑
	150519	150203	0.85	0.94 ↑	0.30	0.89 ↑

are differences in the sampling rate of images or whether additional information (e.g., odometry or information about the current condition) is used or not.

C. Single matching constraints

Even if the algorithms are evaluated on the very same image sets, there are other factors that can have huge influence on the performance measures of an algorithm. Most importantly, this can be the measure itself. For instance, the precision-recall curves in Fig. 4 are recreated according to two research papers ([33] and [25]) of two different groups. Both plots show results of several place recognition algorithms on the exact same Nordland spring-winter dataset.

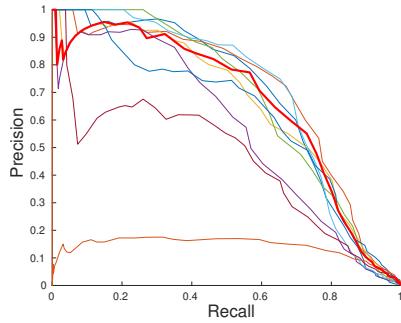


Fig. 3. The performance of an algorithm can vary considerably over different fractions of the same dataset. Each curve shows results of the AlexNet on a different fraction of Nordland spring-winter dataset.

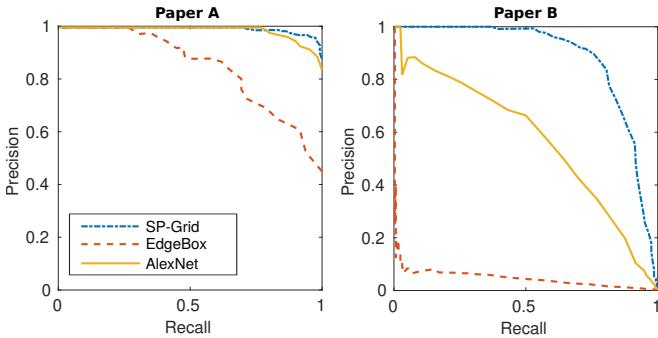


Fig. 4. Two sets of curves on the exact same imagery (Nordland spring-winter) created using evaluation procedures from two different papers (A [25] and B [33]). The same algorithm can show significantly different absolute performance if the evaluation mechanism is changed between papers. Paper A evaluates only the single best matching, paper B evaluates all matchings.

Although some of the algorithms (EdgeBox, Holistic CNN, SP-Grid) are evaluated using the procedure from *both* groups their absolute performance is completely *different*. For example, EdgeBox seems to be a reasonable algorithm for place recognition on this dataset in the right plot, but seems to be unsuited according to the left plot. Although both plots are precision-recall curves of place recognition on the same dataset, the underlying tasks were significantly different. The task evaluated in the left plot was to retrieve *all* relevant database images for a query image, while the task for the right plot was to retrieve a *single* relevant database image (this distinction will also be topic of Sec. IV-C). Obviously, the two tasks are differently difficult and the curves can not be compared *across* tasks (however, at least in this case, the relative ordering of the algorithms within each task is consistent).

D. Details of the usage of the ground truth

Although the usage of the ground-truth matching information is related to the previous more general discussion on dataset details, we want to emphasize the particular importance and problematic nature of ground-truth information for place recognition. Place recognition is an ill-defined problem: In general, the information provided by a query and a database image is not sufficient to definitively answer the question whether they show the same place or not -

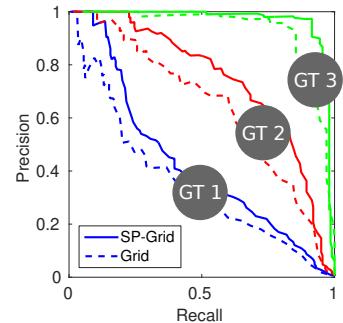


Fig. 5. Influence of a varying maximally allowed distance of images to be considered as ground-truth matchings on the Gardenspoint Walking DayLeft-NightRight sequence. The different ground-truths are indicated by GT1/GT2/GT3. Of course, it makes no sense to compare a solid curve (Grid approach) to a dashed curve (SP-Grid approach) if they use different ground truth, which might be the case if they were from different papers.

this answer always depends on additional information, e.g., a maximally allowed spatial and angular distance between two camera poses.

The here discussed visual place recognition problems are typically in the context of mobile robotics, hand-held cameras, or automotive applications. For images of the same place, we can expect a considerable overlap in their fields of view, but the particular maximum amount of allowed camera pose displacement for images of the same place can be different. Many available datasets come either with associated GNSS (e.g., GPS) information or synchronized image sequences. However, ground truth based on GNSS signals is error prone and requires (manual) correction. Even without such errors, the choice of the maximum spatial and angular distance in a GNSS pose or the maximum distance of frame indexes of ground-truth matchings significantly influences the overall results, an example is illustrated in Fig. 5. Most datasets don't come with an obligatory choice of these thresholds, this allows to tailor them to the level of viewpoint invariance that is suitable for a particular task or algorithm. However, variations also prevent a comparison of results between different papers.

E. Pre- and postprocessing steps

Fig. 1 shows a basic place recognition pipeline with placeholders for descriptor and similarity computation (e.g. simple exhaustive pairwise comparison). This basic pipeline can be extended with pre- and postprocessing steps that can have a large influence on the overall performance. Examples for additional steps are the combined evaluation of multiple sequential images (e.g. [4], [6]) or the standardization of descriptors [3]. The comparison of results for raw and standardized descriptors in table II shows examples of the huge possible improvements due to such an additional simple standardization step; similar results for sequence processing can, e.g., be found in [4].

When comparing the performance of different place recognition approaches, we have to keep in mind that these pre- and postprocessing steps are largely independent of the particular descriptor or comparison algorithm. For instance,

a novel descriptor that already includes this descriptor standardization step has an inherent benefit over other descriptors that do not explicitly integrate standardization. Since descriptor standardization can be seen as a default preprocessing step, the novel descriptor should only be compared against existing descriptors *with* the additional standardization step - otherwise, we would not be able to distinguish whether some improvement is due to the novel descriptor computation or merely due to the (already known) standardization.

F. Limited transitivity of comparison to baseline algorithms

A best practice to compare new algorithms to the state of the art is to compare to (a small set of) repeatedly used baseline algorithms (e.g., FAB-MAP [19] or SeqSLAM [6]). While this provides some information in comparison to the baseline algorithm, we have to be careful to use the baseline algorithms as a link between different papers and experiments. A naive approach could be: If all new algorithms are compared against the same baseline algorithm, we can indirectly compare all new algorithms, e.g., ‘If the existing algorithm X is 10% better than FAB-MAP and my new algorithm Y is 30% better than FAB-MAP, then my new algorithm Y is $\sim 18\%$ better than algorithm X.’.

Of course, in practice, it is not that simple. Although comparison against a baseline provides some information, differences in the evaluation (i.e., those mentioned in this section) prevent this type of math. Moreover, if non-optimized parameter settings are used for the baseline algorithms (e.g., vocabulary in FAB-MAP or the sequence length in SeqSLAM) or datasets are used that are particularly challenging for the baseline algorithms, then it might become very easy to beat the baseline by a large margin but with only little information about the relation to the state of the art. For example, algorithm X could be compared with FAB-MAP on a dataset where FAB-MAP also performs very well (e.g. Oxford [19]), while algorithm Y is able to beat FAB-MAP by a large margin on another dataset that is particularly challenging for FAB-MAP (e.g. Nordland [12]). This is not evidence that Y is better than X on any of the two datasets.

IV. DISTINGUISHING PROPERTIES OF PLACE RECOGNITION PROBLEMS

The previous section demonstrated that we have to be very careful when comparing place recognition results across different experiments. In this section, we will take a closer look at the underlying reasons. We identified seven aspects of visual place recognition experiments whose specification can yield considerably different place recognition problem types with different levels of difficulty. These aspects are listed in table III together with examples for particularly difficult instances.

A. Knowledge level

The amount of available knowledge in a place recognition algorithm can be very different. This is property “A” in table III. In the following, we will discuss four aspects:

(A1) A straightforward aspect is additional sensor information, e.g. from other exteroceptive sensors beyond standard cameras, like stereo, RGB-D, IR or LiDAR, or proprioceptive odometry measures. Additional sensors can considerably simplify the visual place recognition problem. For example, by allowing to measure previously unobservable quantities (e.g. depth with absolute scale) or providing advantages in low-illumination situations (e.g. LiDAR at night). Also related is additional information that can be derived from the (standard) images, but requires additional computational modules, e.g. visual odometry or semantic segmentation [34]. The availability of these modules can create additional limitations, e.g. if no (good) semantic segmentation module for low-illumination conditions in natural environments is available.

(A2) The knowledge level can also vary along the spectrum of online vs. offline problem setups. Along this spectrum, we basically distinguish how much of the database and query set is known in advance. In full offline setups, database and query sets are completely known in advance. On the other end of the spectrum, we can face a full online setup: starting from an empty database, query images are presented once at a time and have to be matched to all previous query images (which then form the database for this query). This is the expected setup in a SLAM system. A practically relevant setup in the middle of this spectrum is a database that is completely known in advance, but the query images are only presented during operating time (this can be expected in many practical localization setups). Setups with previously known database and/or query images allow for extensive preprocessing, e.g. domain specific training [35] or standardization [3].

(A3) Another aspect of the knowledge level is whether there are potentially unknown places during the query run (i.e. places in the query set for which there are no images in the database set). This exploration or open-world property can considerably increase the difficulty to select criteria to establish a matching (i.e. we cannot simply choose the most similar database image).

(A4) In general, any additional knowledge about any of the other here presented aspects can have significant influence on the complexity of the problem. Therefore, it should be obligatory to explicitly list this required knowledge for newly presented approaches. For example, expected levels of viewpoint changes (Sec. IV-B), knowledge about possible appearance conditions (Sec. IV-F), or assumptions about camera trajectories (Sec. IV-G).

B. Amount of viewpoint changes

In contrast to pose estimation, place recognition is an ill-defined problem, since answering the question whether two images show the same place or not requires additional information about the maximum allowed camera pose discrepancy. As said before, an image of the Eiffel tower and Notre-Dame cathedral might be considered showing different places (since there is no common image content) or showing the same place (Paris).

TABLE III
DISTINGUISHING PROPERTIES

Property	(More) simple instances	(More) difficult instances
<i>A - Knowledge level (Sec. IV-A)</i>		
A1 Are there additional sensors?	odometry and/or other sensors	(standard) vision only
A2 online vs. offline	database and query are known in advance	full online
A3 Are there unknown places during query?	for each query image there is a corresponding database image	exploration of potentially unseen places
A4 Is there additional knowledge available about any other aspect?	any additional knowledge or assumption can considerably simplify the problem	no knowledge about potential condition or viewpoint changes, trajectory properties, ...
<i>B - Viewpoint changes (Sec. IV-B)</i>		
B1 amount of viewpoint change	small or no viewpoint changes	large viewpoint changes
B2 Is there a discrete set of places?	well distinguishable discrete places	a continuum of places
B3 degrees of freedom	limited degrees of freedom	full 6-DOF
<i>C - Intended output (Sec. IV-C)</i>		
C1 single best matching or all	single best matching	requesting all matchings
C2 definite matchings or a list of candidates	a high number of potential candidates	definite matchings
C3 similarities or decisions	requesting similarities	requesting hard decisions
<i>D - Dataset scale (Sec. IV-D)</i>		
D1 number of images	small datasets	(very) large scale
D2 runtime requirements	ignoring runtime	hard real-time requirements
D3 storage limitations	ignoring memory consumption	requirement for small descriptors
<i>E - Domain (Sec. IV-E)</i>		
E1 type of environment	well defined and static environment with well distinguishable places	mixed (unknown) and/or dynamic environment with many similar looking places
E2 application area/ vehicle	slow and restricted camera motion	fast and agile camera motion
<i>F - Appearance change (Sec. IV-F)</i>		
F1 character/intensity of changes	(almost) constant appearance of places	severe appearance changes and bad illumination/visibility conditions
F2 Is there a discrete set of condition?	only two or a small number of discrete conditions	continuous conditions
F3 Can the condition change within a sequence?	no in-sequence changes	possibly in-sequence changes
F4 knowledge about conditions	complete knowledge (maybe also including training data)	no knowledge about potential and actual condition
<i>G - Spatio-temporal structure (Sec. IV-G)</i>		
G1 sequences or individual images	long sequences	individual images
G2 assumptions about velocities	constant (or even same) velocity in database and query	arbitrary velocities
G3 Are in-sequence loops possible?	no in-sequence loops	potentially multiple loops
G4 Are stops possible?	no stops	potentially arbitrarily long stops

(B1) The amount of viewpoint change can have significant influence on the difficulty of the place recognition problem and the suitability of a particular approach. For example, changing viewpoints challenge image descriptors that rely on constant positions of image features like the early ConvNet approaches to place recognition that used flattened feature map responses of AlexNet as descriptors [31].

(B2) In another direction, we can distinguish applications and datasets with a discrete set of well separated places in the database (for example where each place is represented by a single image) and those with a continuous sequence of partially overlapping views of an area that is considered to contain several different places. The latter also poses particular challenges to the definition of ground truth (as was illustrated in Fig 5).

(B3) Many mobile robotics place recognition datasets for changing environments contain only relatively small viewpoint changes since they are focused on ground vehicles, often in practical setups with roads or pathways that additionally limit the degrees of freedom of the camera movement. Often there are only longitudinal, small lateral, and small

rotational pose shifts. Datasets with pixel-aligned images like Nordland [12] can be used to systematically evaluate the influence of (synthetic) viewpoint changes.

C. Intended output

A very important distinguishing property is the intended output of a place recognition algorithm. The basic pipeline from Sec. II stated that the output is a set of associations between a query and database sets. However, the characteristics of these associations for a query image can be significantly different. It can be:

- 1) A single best matching.
- 2) (C1) A list of *all* database images that show the same place. This is typically required in online setups when it is not possible to preprocess the database to create a list of unique images per place or when we are interested in multiple matchings, e.g. to add more constraints for the map optimization during SLAM.
- 3) (C2) A list of candidates that will be further refined in a verification step. Here, the intention is to have at least one true-positive matching among the candidates;

- the presence of additional false-positives is irrelevant (which distinguishes this case from the previous).
- 4) (C3) Finally, many algorithms do not output a discrete set of associations but the similarity matrix S . The actual creation of place matchings is then left to further processing stages.

The difficulty can significantly vary between these cases. For example, the OPR [36] approach builds on the assumption of single matches and can benefit from this assumption. However, it is also likely to fail if this assumption is violated in a dataset as shown in [37].

Reducing the output to the similarities instead of discrete matchings brushes the difficulty to make concrete decisions aside, e.g., to select a threshold on the descriptor similarity for a particular dataset. This is also hidden in the evaluation based on (the area under) precision-recall curves that are created based on a variety of thresholds. This also ignores the beneficial property of an algorithm to create large margins between true and false matchings or to allow the usage of the same threshold across many datasets (or procedures to determine the threshold for a dataset).

D. Dataset scale and runtime/memory requirements

(D1) The number of expected database and query images can significantly vary. The datasets that are typically used for evaluation in mobile robotics range from a few hundred images (e.g. GardensPoint Walking [10]) to several thousand (e.g. Oxford Robot Car [13]) and typically cover rather small areas like a campus or a few roads of a city.

(D2, D3) Dependent on the practical application context of place recognition there can be limitations on the available time for each query (including real-time requirements) and limitations on the available storage space for the database. As discussed in Sec. II many place recognition approaches combine a descriptor computation stage with a matching stage. For small scale datasets, the computational costs at the descriptor stage might dominate the overall computation costs. However, the effort for computing descriptors depends linearly on the numbers n and m of images in the database and query sets ($O(n + m)$) but the costs for an exhaustive pairwise comparison grows significantly faster ($O(n \cdot m)$). Therefore, (very) large-scale place recognition problems pose significantly different challenges not only for descriptiveness but also regarding runtime and memory consumption [38], [37].

E. Domain

The application domain has very large influence on many properties from table III. Thus there is some overlap of this section with others.

(E1) One important aspect of the domain is the type of environment for place recognition. Typical distinctions are indoor-outdoor or urban-rural-natural (see Fig. 6 for some example images). Some environments tend to have different places that look very similar. This visual aliasing is particularly challenging for place recognition. The environment also has important influence on the expected type and amounts of



Fig. 6. Example images for the most common types of environment in large-scale place recognition

appearance changes, viewpoint change, and dynamic objects. It is also particularly important when integrating additional semantic information (e.g. recognizing the type of a room) or when using learning-based methods like NetVLAD [32] or DELF [39] that were trained on datasets with images from one type of environment and could perform worse if applied in environments different to the training data.

(E2) Another important aspect of the domain is the application area and, strongly related, the device/vehicle that provides the (query) images. For example, place recognition in an automotive scenario exhibits less viewpoint change but more dynamic objects than an autonomous air vehicle that flies through a forest. The application area is also expected to influence the criticality of false-positive matchings and the intended output as described before in Sec. IV-C.

F. Appearance changes

Long-term operation of robots and long-term usage of vision-sensor-based maps create an increasing interest in place recognition despite severe appearance changes of the environments, e.g. due to lighting (e.g., dawn, sunny, dusk, night), weather (e.g., sun, rain, snow, storm), or seasonal changes (e.g., summer/winter, dry/wet season). Of course, the intensity and type of appearance change has significant influence on the difficulty of a place recognition problem setup and on the selection of suitable approaches. However, there are also other, less obvious aspects with potentially significant influence. Overall, we identified the following four aspects of appearance changes as distinguishing properties of place recognition problems:

(F1) *Character and intensity of appearance change*: The spectrum ranges from mild lighting changes to day-night or winter-summer changes. Often, there is a trade-off between robustness towards appearance and viewpoint changes. Methods like FAB-MAP [19] have shown good performance for visual place recognition in environments with dynamic objects and slight changes in illumination. In contrast, visual place recognition in *changing* environments is still challenging and therefore subject of active research. For example, the authors of [40] observed in their localization experiments that local SURF features [41] fail under strong lighting changes. Similarly, [33] concluded that the performance of the SIFT-DoG-6 detector for place recognition drops with increasing severity of condition changes.

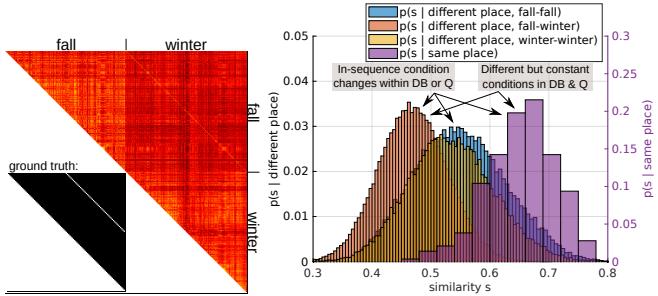


Fig. 7. Descriptor similarities s from pairwise AlexNet descriptor [31] comparisons S and corresponding conditional probabilities $p(s|\cdot)$ from the Nordland dataset with in-sequence condition change from fall to winter. The ground-truth matchings (see bottom left) in fall-winter cannot be well distinguished from *different* places in winter-winter and fall-fall.

(F2) Is there a discrete set of conditions or a continuum?

Most place recognition approaches for changing environments are evaluated on datasets where the database and query sequence have different conditions, but each has a single condition out of a discrete set of possible conditions. This can be exploited, for example, in domain adaptation approaches that are able to translate images between conditions (e.g., *ToDayGAN* [42]) or by simple standardization approaches (see [3] for a discussion). However, in many real long-term operation scenarios, we can expect a continuum of appearance conditions, for example, various types of “rainy”, a smooth transition between day and night, and no clear distinction between a blue sky and overcast conditions. Please refer to [3] for a more detailed discussion about continuous condition changes.

(F3) Can the condition (continuously) change within the database and/or query set? Strongly related to the previous aspect is the question whether conditions can change *within* the database or query sequence. As recently discussed in [3], most approaches and datasets in the literature assume different but constant conditions within database and query, for example database is “day” and query is “night”. But what if the query sequence was captured during the transition from day to night? Fig. 7 uses the example of place recognition on Nordland using a combined “fall-winter” condition to illustrate the problem for the choice of the matching criteria: Different places under the same condition tend to look more similar than the same place under different conditions.

Fig. 7 shows the corresponding pairwise image similarity matrix S obtained from the AlexNet-conv3 descriptor [31] with ground truth, and conditional probabilities $p(s|\cdot)$ for all similarities s in S . Since we assume a single online run and compare each current image only to its predecessors, an upper triangular matrix is received. After an initial traverse in fall without loops, the camera revisits all places once in winter. Accordingly, high similarities (i.e. bright pixels) should only appear as single minor diagonal in *fall-winter* image comparisons. However, S (Fig. 7, left) has high similarities (i.e. bright pixels) as well in *fall-fall* and *winter-winter* image comparisons, although these do not contain similarities of same places.

The histograms in Fig. 7 (right) support this intuition: In place recognition experiments with different but constant conditions in database and query, we would merely observe the most left probabilities $p(s|\text{different place, fall-winter})$ and most right probabilities $p(s|\text{same place})$. These two distributions have only a small overlap and are therefore well separable which leads to a high place recognition performance. In contrast, in continuously changing environments we additionally obtain the probabilities $p(s|\text{different place, fall-fall})$ and $p(s|\text{different place, winter-winter})$ that have a much higher overlap with $p(s|\text{same place})$ in Fig. 7 (right). As a consequence, same and different places are less separable which can, if not taken care of, lead to a significant performance drop.

(F4) Knowledge about possible conditions: Finally, for each of the above aspects, the level of knowledge is crucial. This includes knowledge about the principle problem setup (Which conditions might appear?), knowledge about the particular problem instance (What is the current condition?) as well as the availability of training data, e.g. to learn to detect the current condition, learn invariant representations, or to translate between different conditions.

G. Spatio-temporal structure

(G1) In contrast to a general image retrieval problem, for the visual place recognition problem considered here, images are typically recorded by a robot, car, or person equipped with a camera that is driving or walking along a trajectory in an environment. Consequently, images in database and query might be acquired as spatio-temporal *sequences*. Consecutive images show either same or neighbouring places, and were likely recorded under the same or a very similar condition. If we then compare images from a database and query, we likely receive a similarity matrix S with one or multiple salient sequences of high similarities like in Fig. 8 (left). This is exploited by sequence-based place recognition approaches to improve the place recognition performance [6], [4], [44], [45].

However, dependent on the trajectories, velocities, and frame rates in the database and query sequences, the spatio-temporal structure (and thus the sequences in the similarity and ground-truth matrices) can significantly vary. Typical patterns are illustrated in Fig. 9:

- 1) If the camera moves non-stop and never revisits a same place within database and query (but there are revisits between database and query), a single sequence similar to Fig. 8 (left) can be observed.
- 2) If places are revisited *within* database or query (*loops*), additional sequences with high similarities s in S appear (Fig. 9, left).
- 3) If the camera temporarily *stops* or moves very slowly on its trajectory, consecutive images show the same place in database or query. If the camera stops merely in the database and moves at the same place in the query, a vertical sequence of high similarities appears in S (Fig. 9, right). Similarly, if the camera stops

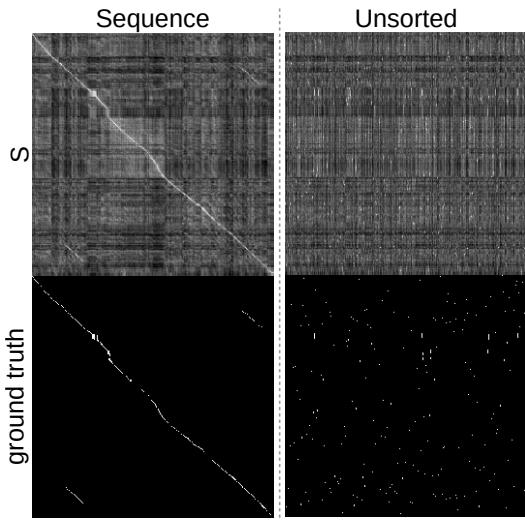


Fig. 8. It is much easier to find all matchings in the similarity matrix S of a temporally ordered sequence than in the unordered image set (part of StLucia 08:45).

only in the query and moves at the same place in the database, a horizontal line of high similarities appears in S (Fig. 9, right). And finally, if the camera stops at the same place in both database and query, a rectangle of high similarities in S can be observed (Fig. 9, right).

- 4) If new places are explored in the query, low similarities should be observed for all explored query images over all database images (Fig. 9, left).

This is important since many sequence-based place recognition approaches make (implicit) assumptions about this structure and potentially fail if these assumptions are not met. Dependent on whether these assumption can be enforced by appropriate preprocessing or not (e.g. using visual odometry and/or image resampling), they might limit the general applicability of a place recognition approach. Typical assumptions about spatio-temporal structure are:

- 1) (G2) Constant velocities in the database and query (or even the same velocity)
- 2) (G3) No loops within the database sequence (and potentially also the query).
- 3) (G4) No stops in database and/or query.
- 4) (cf. A3) No unseen places during query.

V. DISCUSSION

The previous sections argued that place recognition experiments are barely comparable across papers and that there is a variety of properties that can change from one experiment to another. We think this can have severe implications for different involved parties: people who just want to use place recognition as a tool, researchers that look for open research questions, authors that want to create good papers on this topic, and also reviewers that have to identify problems in submitted papers.

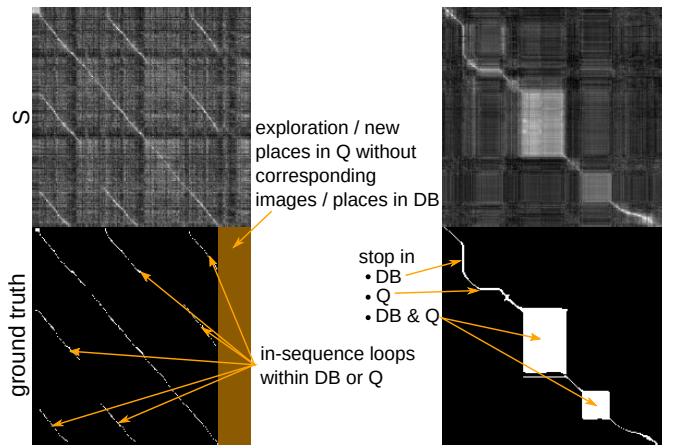


Fig. 9. Typical patterns in pairwise similarity matrices $S \in \mathbb{R}^{|DB| \times |Q|}$ between the database (DB) and query (Q) set that are caused by the spatio-temporal structure of datasets.

A. There is no one-fits-all solution (yet?)

The large variety of properties of place recognition problems makes it very hard to design a single approach that meets all potential requirements. The ill-defined nature of the place recognition problem can create incompatible requirements. For example, one application might call for image matchings despite severe viewpoint changes, another might be much more restrictive on possible deviations of the camera pose. From a more general perspective, as in many problems, the more specific we can tailor the solution to the particular problem instance, the better it can solve the problem. Often, this tailoring is based on the introduction of additional assumptions on the previously discussed properties. For example, assumptions about possible camera motions (Sec. IV-G) or potential appearance changes (Sec. IV-F). Introducing such bias in the algorithm can make it completely unsuited for problem instances where the assumptions are violated.

A promising approach to address this issue is to automatically determine the properties of the place recognition problem at hand and to select the appropriate approach. However, this seems to be largely unexplored terrain for place recognition.

B. There are open research questions

Although place recognition is an extensively studied problem, there are challenging open research questions. The sheer number of properties that can vary in the setup of a place recognition experiment makes the creation of a standard benchmark an important and challenging research task on its own.

The third column of table III provides examples of (more) difficult instances of each property. In particular, problem setups, that combine multiple of these difficult properties are also likely to be open research topics. For example, a A1+B1+F1+G1-problem that is (standard) camera only (A1), with large viewpoint changes (B1), potentially severe appearance changes (F1), and no spatio-temporal sequences (G1).

Examples for individual properties that challenge existing approaches are in-sequence appearance changes (F3) and finding *all* database matchings for a query instead of a single match (C1). Of course, very bad illumination or visibility conditions (F1, heavy rain, snow, night-time, overexposure, etc.) are a challenge on their own and improvements on local and holistic image descriptors can still make a huge difference.

With a more general perspective, computational demands (D1+D2+D3) and power consumption can be critical on mobile devices like robots. For example, their limited resources can hamper the usage of powerful yet expensive approaches like pairwise local feature matching and usage of epipolar geometry for more than a few candidates per query. Resource-efficient usage of the representative power of local features in combination with geometric verification can also be a promising direction for future research.

As said before, based on the observation that additional assumptions can considerably simplify the problem and their violation can bungle a particular approach, research towards automatic testing of the validity of such assumptions can potentially have huge impact for practical applicability.

C. Papers have to provide a lot of information about the problem and experimental setups

As long as there is no such automatic adaptation, for practical application it is crucial to know the properties of the place recognition problem at hand, as well as the properties, assumptions, and limitations of the existing algorithmic approaches. Basically, each paper that introduces a novel place recognition approach and/or provides experimental results should provide information about each of the properties from table III. This information is also essential to ensure repeatability of experiments and to increase the comparability between papers.

D. Cherry picking is easy

The previous Sec. IV showed the variety of different place recognition problem types. It is clear that different place recognition approaches might be differently well suited for different problem types. For example, one image description approach might be more robust to appearance changes (e.g., AlexNet) and another more robust towards viewpoint changes (e.g., SIFT keypoints). Table II demonstrated that the relative performance of NetVLAD and AlexNet can switch dependent on the evaluated dataset. Given the large variety of existing datasets (cf. table I), there is a good chance that for some new algorithms there is one or a few datasets where it can beat some other algorithms.

An important observation is that such different strengths of approaches are often also reflected in the design of experiments and evaluations of papers that propose new approaches. However, this can potentially hide weaknesses of the approach or requirements for its applicability.

For sequence-based approaches, an often made (and sometimes hidden) simplification are restrictions on the trajectory of the camera, like assuming constant velocity between

database and query sequences. Even if the origin velocities of the cameras might vary, some papers apply a resampling of the camera video stream to discrete images based on the ground-truth poses. This resampling can simplify the problem to merely searching for main or secondary diagonals in the pairwise image similarity matrix. Of course, the ground-truth pose knowledge is not available in practical applications which makes this approach flawed.

Basically, all (more) simple instances of properties from table III can be used to simplify the problem. For example, assuming that there are no unseen places during query (A3), aligned viewpoints (B1), only a single matching database image for each query (C1), knowledge about possible conditions (F4), and so on.

To make this clear, there is nothing wrong in making those assumptions, as long as (1) they are clearly stated, (2) they are realistic in some (practical) scenario, and (3) the assumption is appropriately reflected in the experimental comparison to other approaches, in particular by comparing against methods that also exploit this simplification (if there are any).

An example for a violation of the last point was discussed in Sec. III-E: proposing a new descriptor for place recognition that includes a descriptor standardization step and then comparing against existing approaches without this standardization.

To reduce or prevent confusion about assumptions and properties, all additional assumptions, required knowledge, or pre/postprocessing steps should be made clear and justified for newly presented place recognition approaches. Moreover, evaluation should always be performed on a broad range of datasets, e.g. 10-20 sequence comparisons from multiple datasets. An indicator for suspect cherry picking can be when the selection of evaluated datasets changes from one experiment to the other within one paper without explanation.

VI. CONCLUSIONS

We discussed that place recognition results are barely comparable across papers and that there is a variety of properties that can lead to considerably different types of place recognition problems. We hope this increases the awareness of these aspects and supports according treatment in future research. The list of properties is not necessarily complete (nor is the chosen partitioning obligatory) and is likely to grow with future developments and further specialization in this field. If appropriate, we will consider an update of this list and any comments and further contributions are very much welcome.

REFERENCES

- [1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016.
- [2] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021.
- [3] S. Schubert, P. Neubert, and P. Protzel. Unsupervised learning methods for visual place recognition in discretely and continuously changing environments. In *Intl. Conf. on Robotics and Automation (ICRA)*, 2020.

- [4] P. Neubert, S. Schubert, and P. Protzel. A neurologically inspired sequence processing model for mobile robot place recognition. *IEEE Robotics and Automation Letters*, 4(4):3200–3207, Oct 2019.
- [5] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *International Conference on Computer Vision (ICCV)*, 2015.
- [6] M. J. Milford and G. F. Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *International Conference on Robotics and Automation (ICRA)*, 2012.
- [7] H. Badino, D. Huber, and T. Kanade. Visual topometric localization. In *Intelligent Vehicles Symposium (IV)*, 2011.
- [8] T. Naseer, W. Burgard, and C. Stachniss. Robust visual localization across seasons. *IEEE Transactions on Robotics*, 34(2):289–302, 2018.
- [9] Freiburg and Bonn Visual Place Recognition Datasets. <http://www.ipb.uni-bonn.de/data/visual-place-recognition-datasets/>.
- [10] Arren Glover. Day and night with lateral pose change datasets. 2014.
- [11] Nicholas Carlevaris-Bianco, Arash K Ushani, and Ryan M Eustice. University of michigan north campus long-term vision and lidar dataset. *The International Journal of Robotics Research*, 35(9):1023–1035, 2016.
- [12] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. *International Conference on Robotics and Automation (ICRA) Workshop on Long-Term Autonomy*, 2013.
- [13] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.
- [14] Jake Bruce, Jens Wawerla, and Richard Vaughan. The SFU mountain dataset: Semi-structured woodland trails under changing environmental conditions. In *International Conference on Robotics and Automation (ICRA) Workshop on Visual Place Recognition in Changing Environments*, 2015.
- [15] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth. Fab-map + ratslam: Appearance-based slam for multiple times of day. In *International Conference on Robotics and Automation (ICRA)*, 2010.
- [16] Shane Griffith, Georges Chahine, and Cédric Pradalier. Symphony lake dataset. *The International Journal of Robotics Research*, 36(11):1151–1158, 2017.
- [17] F. Maffra, L. Teixeira, Z. Chen, and M. Chli. Loop-closure detection in urban scenes for autonomous robot navigation. In *International Conference on 3D Vision (3DV)*, 2017.
- [18] Yukung Choi, Namil Kim, Kibaek Park, Soonmin Hwang, Jae Shin Yoon, and In So Kweon. All-day visual place recognition: Benchmark dataset and baseline. In *Computer Vision and Pattern Recognition (CVPR) Workshop on Visual Place Recognition in Changing Environments*, 2015.
- [19] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.
- [20] H. Lategahn, J. Beck, B. Kitt, and C. Stiller. How to learn an illumination robust image feature for place recognition. In *Intelligent Vehicles Symposium (IV)*, 2013.
- [21] Gaurav Pandey, James R McBride, and Ryan M Eustice. Ford campus vision and lidar data set. *The International Journal of Robotics Research*, 30(13):1543–1552, 2011.
- [22] Sourav Garg and Michael Milford. Straightening sequence-search for appearance-invariant placerecognition using robust motion estimation. In *Australasian Conference on Robotics and Automation (ACRA)*, 2017.
- [23] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [24] José-Luis Blanco-Claraco, Francisco Ángel Moreno-Dueñas, and Javier González-Jiménez. The málaga urban dataset: High-rate stereo and lidar in a realistic urban scenario. *The International Journal of Robotics Research*, 33(2):207–214, 2014.
- [25] Niko Sünderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. In *Proceedings of Robotics: Science and Systems*, 2015.
- [26] C. Valgren and A. J. Lilienthal. Sift, surf and seasons: Long-term outdoor localization using local features. In *European Conference on Mobile Robots (ECMR)*, 2007.
- [27] A. R. Zamir and M. Shah. Image geo-localization based on multipleinarest neighbor feature matching usinggeneralized graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1546–1558, 2014.
- [28] C. Cadena and J. Neira. A learning algorithm for place recognition. In *International Conference on Robotics and Automation (ICRA) Workshop on Long-term Autonomy*, 2011.
- [29] Michael Warren, David McKinnon, Hu He, and Ben Upcroft. Unaided stereo vision based pose estimation. In *Australasian Conference on Robotics and Automation (ACRA)*, 2010.
- [30] Edward Pepperell, Peter Corke, and Michael Milford. Towards persistent visual navigation using smart. In *Australasian Conference on Robotics and Automation (ACRA)*, 2013.
- [31] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford. On the performance of convnet features for place recognition. In *International Conference on Intelligent Robots and Systems (IROS)*, 2015.
- [32] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *Trans. on Pattern Analysis and Machine Intelligence*, 40(6), 2018.
- [33] Peer Neubert. Superpixels and their Application for Visual Place Recognition in Changing Environments. PhD thesis, Chemnitz University of Technology, 2015.
- [34] Peer Neubert, Stefan Schubert, Kenny Schlegel, and Peter Protzel. Vector semantic representations as descriptors for visual place recognition. In *Proc. of Robotics: Science and Systems (RSS)*, 2021.
- [35] C. McManus, B. Upcroft, and P. Newman. Scene signatures: Localised and point-less features for localisation. In *Robotics: Science and Systems*, 2014.
- [36] O. Vysotska and C. Stachniss. Lazy data association for image sequences matching under substantial appearance changes. *IEEE Robotics and Automation Letters (RA-L)*, 1(1):213–220, 2016.
- [37] Stefan Schubert, Peer Neubert, and Peter Protzel. Beyond ANN: Exploiting structural knowledge for efficient place recognition. In *Intl. Conf. on Robotics and Automation (ICRA)*, 2021.
- [38] Peer Neubert and Stefan Schubert. Hyperdimensional computing as a framework for systematic aggregation of image descriptors. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [39] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *International Conference on Computer Vision (ICCV)*, 2017.
- [40] Paul Furgale and Timothy D. Barfoot. Visual teach and repeat for long-range rover autonomy. *Journal of Field Robotics*, 27(5):534–560, 2010.
- [41] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, 2006.
- [42] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. V. Gool. Night-to-day image translation for retrieval-based localization. In *International Conference on Robotics and Automation (ICRA)*, 2019.
- [43] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman. Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles. In *International Conference on Robotics and Automation (ICRA) Workshop on Visual Place Recognition in Changing Environments*, 2014.
- [44] Tayyab Naseer, Luciano Spinello, Wolfram Burgard, and Cyrill Stachniss. Robust visual robot localization across seasons using network flows. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [45] Stefan Schubert, Peer Neubert, and Peter Protzel. Fast and memory efficient graph optimization via icm for visual place recognition. In *Proc. of Robotics: Science and Systems (RSS)*, 2021.
- [46] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*, 1999.
- [47] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *International Conference on Computer Vision (ICCV)*, 2011.
- [48] R. Mur-Artal and J. D. Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgbd cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.
- [49] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

- [50] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [51] C.L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision (ECCV)*, 2014.
- [52] Peer Neubert and Peter Protzel. Beyond holistic descriptors, keypoints, and fixed patches: Multiscale superpixel grids for place recognition in changing environments. *IEEE Robotics Autom. Lett.*, 1(1):484–491, 2016.
- [53] S. M. Lowry, M. J. Milford, and G. F. Wyeth. Transforming morning to afternoon using linear regression techniques. In *International Conference on Robotics and Automation (ICRA)*, 2014.
- [54] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *European Conference on Computer Vision (ECCV)*, 2010.
- [55] N. Sünderhauf and P. Protzel. Brief-gist - closing the loop by simple means. In *International Conference on Intelligent Robots and Systems (IROS)*, 2011.
- [56] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [57] R. Arandjelovic and A. Zisserman. All about vlad. In *Conf. on Computer Vision and Pattern Recognition*, 2013.
- [58] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.
- [59] Bingyi Cao, André Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European Conference on Computer Vision (ECCV)*, 2020.
- [60] P. Neubert and P. Protzel. Local region detector + cnn based landmarks for practical place recognition in changing environments. In *European Conference on Mobile Robots (ECMR)*, 2015.
- [61] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004.
- [62] D. Galvez-López and J. D. Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.
- [63] W. Li, Y. Zhang, Y. Sun, W. Wang, M. Li, W. Zhang, and X. Lin. Approximate nearest neighbor search on high dimensional data — experiments, analyses, and improvement. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1475–1488, 2020.
- [64] Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020.
- [65] P. Neubert, S. Schubert, and P. Protzel. Exploiting intra database similarities for selection of place recognition candidates in changing environments. In *Computer Vision and Pattern Recognition (CVPR) Workshop on Visual Place Recognition in Changing Environments*, 2015.
- [66] O. Vysotska, T. Naseer, L. Spinello, W. Burgard, and C. Stachniss. Efficient and effective matching of image sequences under substantial appearance changes exploiting gps priors. In *International Conference on Robotics and Automation (ICRA)*, 2015.
- [67] E. Pepperell, P. I. Corke, and M. J. Milford. All-environment visual place recognition with smart. In *International Conference on Robotics and Automation (ICRA)*, 2014.
- [68] S. Schubert, P. Neubert, and P. Protzel. Towards combining a neocortex model with entorhinal grid cells for mobile robot localization. In *European Conference on Mobile Robots (ECMR)*, 2019.
- [69] J. M. Á. Alvarez and A. M. Lopez. Road detection based on illuminant invariance. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):184–193, March 2011.
- [70] A. Ranganathan, S. Matsumoto, and D. Ilstrup. Towards illumination invariance for visual localization. In *Int. Conf. on Robotics and Automation (ICRA)*, 2013.
- [71] P. Corke, R. Paul, W. Churchill, and P. Newman. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In *Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2013.
- [72] M. Shakeri and H. Zhang. Illumination invariant representation of natural images for visual place recognition. In *Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2016.
- [73] Zhenqiang Ying, Ge Li, Xianghao Zang, Ronggang Wang, and Wenmin Wang. A novel shadow-free feature extractor for real-time road detection. In *ACM Intl. Conf. on Multimedia*, 2016.
- [74] P. Neubert, N. Sünderhauf, and P. Protzel. Appearance change prediction for long-term navigation across seasons. In *European Conf. on Mobile Robots (ECMR)*, 2013.
- [75] S. Lowry and M. J. Milford. Supervised and unsupervised linear learning techniques for visual place recognition in changing environments. *IEEE Transactions on Robotics*, 32(3):600–613, June 2016.
- [76] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool. ComboGAN: Unrestrained scalability for image domain translation. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.
- [77] S. Garg, B. Harwood, G. Anand, and M. Milford. Delta descriptors: Change-based place representation for robust visual localization. *IEEE Robotics and Automation Letters*, 5(4):5120–5127, 2020.
- [78] 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 *Int. Conf. on Rob. & Autom.*, 2015.
- [79] O. Vysotska and C. Stachniss. Relocalization under substantial appearance changes using hashing. *Int. Conf. on Intelligent Robots and Systems Workshop PPNIV’17*, 2017.
- [80] Yang Liu and Hong Zhang. Visual loop closure detection with a compact image descriptor. In *Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2012.
- [81] Jesse Davis and Mark Goardrich. The relationship between precision-recall and roc curves. In *International Conference on Machine Learning (ICML)*, 2006.
- [82] Claude Sammut and Geoffrey I. Webb, editors. *Encyclopedia of Machine Learning*. Springer US, Boston, MA, 2010.
- [83] B. Ferrarini, M. Waheed, S. Waheed, S. Ehsan, M. J. Milford, and K. D. McDonald-Maier. Exploring performance bounds of visual place recognition using extended precision. *IEEE Robotics and Automation Letters*, 5(2):1688–1695, 2020.
- [84] N. Sünderhauf and P. Protzel. Switchable constraints for robust pose graph slam. In *International Conference on Intelligent Robots and Systems (IROS)*, 2012.
- [85] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla. Benchmarking 6dof outdoor visual localization in changing conditions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
- [86] Huanhuan Fan, Yuhao Zhou, Ang Li, Shuang Gao, Jijunnan Li, and Yandong Guo. Visual localization using semantic segmentation and depth prediction. 2020.
- [87] Jianzhu Huai, Yusen Qin, Fumin Pang, and Zichong Chen. Segway drive benchmark: Place recognition and slam data collected by a fleet of delivery robots. *arXiv*, 2019.
- [88] A. Pronobis, B. Caputo, P. Jensfelt, and H.I. Christensen. A realistic benchmark for visual indoor place recognition. *Robotics and Autonomous Systems*, 58(1):81 – 96, 2010.

APPENDIX A: DETAILS ON THE BASIC PLACE RECOGNITION PIPELINE

A. Input

Input to a place recognition pipeline is a set DB of n database images and a set Q of m query images. Depending on the application, DB and Q can be separate (e.g. global localization) or identical (e.g. SLAM). In an *offline* setup, all images in Q are given in advance; in an *online* setup, Q is a growing set of images; a setup for *delayed* place recognition uses a few future query images for performance improvements. For *realtime* requirements, new query images have to be processed in a certain amount of time. Q can solely contain images of same places as in DB or also images of additional, unseen places (*exploration*).

B. Descriptor Computation

The first step in the basic place recognition pipeline is the computation of some kind of descriptor from each database and query image. Existing descriptors can generally be divided into local and holistic descriptors, which can both be either hand-designed or learned. Hand-designed local descriptors like SIFT [46], SURF [41], or ORB [47] detect and describe local features in an image, and demonstrated good performance for place recognition in SLAM systems like FAB-MAP [19] or ORB-SLAM2 [48]. However, it was shown several times in the literature that existing local feature detectors fail under environmental condition changes [26], [40], [49], [33, p.187]. Therefore, deep-learned local descriptors like DELF [39] or D2-Net [50] were proposed that achieve good performance despite condition changes. There are also combined systems like EdgeBoxes [51] or SP-Grid [52]. The drawback of local descriptors is the high computational effort when comparing two images for place recognition.

An often faster alternative are holistic descriptors that describe an image as a single (high-dimensional) vector. In earlier work on place recognition, preprocessed pixel-values directly served as holistic descriptor [6], [53] or a single hand-designed descriptor for local features like BRIEF [54] was computed from a full image [55]. Later, the work [49] proposed DenseVLAD which computes local RootSIFT descriptors [56] in a grid all over an image and combines them via VLAD [57] in a single holistic image descriptor. The authors of [31] proposed the usage of intermediate layer activations from Convolutional Neural Networks (CNNs) trained for object classification like AlexNet [58]. Descriptors like NetVLAD [32] or DELG [59] are CNNs directly trained for place recognition tasks.

C. Computing the similarity between two images

Given holistic or local descriptors, there are different ways to compute the similarity (or distance) between two images. The mean absolute error [6], [30] or mean squared error [53] can be used to compare the preprocessed pixel-values of two images. Cosine similarity (e.g., [31]) or euclidean distance (e.g., [32]) is mainly used to compare holistic descriptors.

For local image descriptors, the similarity computation between two images is often divided into two steps: 1) find and validate local descriptor matches, 2) compute the image similarity based on the set of local features. In the first step, local feature correspondences between two images can be found for example with a left-right check (also termed crosscheck or mutual matchings) [60, Eq.4] or with a geometric verification [39], [33]. Additional aspects like the ratio between the closest and the next closest descriptors [61, p.104] have been used as well.

Given the descriptor correspondences, the similarity can be computed in different ways. In [60, Eq.4] the average similarity of all matching pairs is computed; [25] adds a term for the local feature's box shape. In [39], the number of inliers after geometric verification is directly used as similarity score.

D. Which images to compare between database and query

Another important aspect that affects runtime and performance is which database images are compared to a current query. Many approaches simply compare all images from the database with all query images (e.g., [6], [44]). An alternative approach is to use a nearest neighbor search [50] with methods like Bags of Words [62] or different approximate nearest neighbor methods [63] like HNSW [64]. A third way to reduce the number of image descriptor comparisons is the exploitation of spatio-temporal sequences within database and query (see Sec. IV-G) by selecting matching candidates from the database for the current query from the previous matching pairs [36], [65], [37]. In [66], available noisy GPS was used for candidate selection.

E. Extensions of the basic place recognition pipeline

There is a variety of existing extensions to the basic pipeline for performance improvements that build upon different assumptions. The following methods from the literature are roughly sorted from input to output within the basic pipeline (Fig. 1). 1) Additional sensor data: If available, additional sensor data like odometry [67], [68] can be used. 2) Image preprocessing: Some methods in the literature exploit prior knowledge about the condition changes. If only the lighting condition changes due to time of day, illumination-invariant image conversion [69], [70], [43] or shadow removal [71], [72], [73] can be used. In case of different but known condition changes, the conversion of query images into the database condition with appearance change prediction [74], linear regression [75] or Generative Adversarial Neural Networks (GANs) [76], [42] have been proposed. 3) Descriptor preprocessing: Particularly in an offline setup with all database and query images, a separate feature standardization of all descriptors for each condition [3] can be used to clearly improve the performance. Without knowledge about when the conditions change, unsupervised learning based on clustering can be used [3]. 4) Sequence-based methods: Since place recognition is an embedding of image retrieval into mobile robotics, spatio-temporal sequences within database and query can often be exploited. MCN [4] and Delta Descriptors [77] encode multiple consecutive descriptors into a new descriptor for each image. Sequence-based methods like SeqSLAM [6], ABLE [78], HMM and methods based on Flow Networks [44], [79] postprocess the similarity matrix $S \in \mathbb{R}^{|DB| \times |Q|}$. 5) General postprocessing of S : Unsupervised learning methods based on PCA [80], [75] have been used to improve the pairwise image similarities [3].

F. Place recognition evaluation metrics

Place recognition evaluation typically builds upon knowledge of ground-truth place associations. For most available datasets this information is either given as a list of image pairs from DB and Q that show the same place or by an associated pose (e.g. based on GNSS) for each image. Thresholds on the (lateral, longitudinal, and rotational) distance between poses can then be used to compute a list of

ground-truth matchings between DB and Q . The assumed binary nature of this ground-truth information is an essential property of the ill-defined place recognition problem. This and the importance of the actual choice of the above pose thresholds is discussed in Sec. III-D. The binary matrix $GT \in \{0, 1\}^{n \times m}$ of the same size as S indicates whether a pair of database and query image show the same place or not.

An often used basic evaluation pipeline is to take the similarity matrix S as the output of the place recognition approach and to apply a series of varying thresholds t_k with

$$t_k = [\min(S), \min(S) + \epsilon, \min(S) + 2\epsilon, \dots, \max(S)] \quad (1)$$

on S in order to obtain binary matching decisions.

Although the actual choice of the threshold is practically very important, this allows to evaluate the quality of the similarities independent of the quality of a particular threshold. A low threshold t_k is likely to create many image pairs which are actually correct (true positives TP_k) but also potentially classifies many image pairs as matchings which actually do not show the same place (false positives FP_k). A high threshold in turn helps to prevent false positives but increases the risk of also missing database images that show the same place as the query (false negatives FN_k). Given the binary matching decisions, the ground truth can be used to count these numbers [81]:

$$TP_k = |(S \geq t_k) \wedge GT| \quad (2)$$

$$FP_k = |(S \geq t_k) \wedge \neg GT| \quad (3)$$

$$FN_k = |(S < t_k) \wedge GT| \quad (4)$$

These values can then be used to compute the k -th point on a precision-recall-curve (PR-curve) by

$$\text{Precision: } P_k = TP_k / (TP_k + FP_k) \quad (5)$$

$$\text{Recall: } R_k = TP_k / (TP_k + FN_k) \quad (6)$$

Precision-recall-curves are finally defined as the precision as a function of the recall with

$$P_k = f(R_k) \quad (7)$$

and serve as a widely used performance metric in the literature on place recognition. Note that ROC-curves are a related performance measure but are unsuited due to the imbalance between matching and non-matching places [81]. While precision-recall curves provide detailed information about the performance of an algorithm, they require much space for visualization. Therefore, they are not suited very well for an evaluation over many datasets.

To compress the information into a single number, different approaches have been used:

- The area under the precision-recall curve (AUC)

$$AUC = \int_0^1 P(R) dR \quad (8)$$

- Maximum F1 score, which is the maximum harmonic mean of P_k and R_k [82, p.781] with

$$\text{maximum F1 score} = \max_k \frac{2 \cdot P_k \cdot R_k}{P_k + R_k} \quad (9)$$

While AUC contains information about the full precision-recall curve, the maximum F1 score merely represents a single point on the precision-recall curve that represents a potentially good compromise between high precision and high recall.

- Maximum recall at 100% precision (Recall@100%Precision). This has later been modified to Extended Precision [83] to address the problem of Recall@100%Precision, if the precision never hits $P = 100\%$. The Recall@100%Precision was invented, because 100% precision ensures solely correct loop closure in SLAM. However, since the advent of robust pose graph optimization techniques around 2012 [84], avoiding wrong loop closures became less important.

An somewhat complementary metric is Recall@K which counts for which proportion of query images with true-matchings, one of the true-matchings is actually in the K highest rated database images. It expresses the performance of a place recognition pipeline for candidate selection for global localization.

In addition to evaluation of place recognition performance, runtime and memory consumption are likely practically very important aspects.

APPENDIX B: EXISTING PLACE RECOGNITION BENCHMARKS

1) The Visual Localization Benchmark [85]: This benchmark¹ was designed to evaluate and compare approaches for visual localization in changing environments. The task is to compute for each query image a transformation estimation (x, y, θ) that is then compared to three maximum-pose-error thresholds *hard* (e.g., $(0.25m, 2^\circ)$), *middle* (e.g., $(0.5m, 5^\circ)$), and *easy* (e.g., $(5m, 10^\circ)$). Their benchmark metric is described as: “*the percentage of query images which were localized within three given translation and rotation thresholds*”. The used datasets with the roughest threshold could potentially be used for place recognition.

Approaches that show very good visual localization performance use place recognition techniques for candidate selection before conducting expensive transformation estimation. For example the winner of the *CVPR’20 Localization Challenge*² used NetVLAD [32] for visual place recognition and candidate selection before feature computation and matching³. Other methods that performed very well on this benchmark like D2-Net [50] or ONavi [86] used NetVLAD [32] or DELF [39], respectively, for place recognition.

¹The Visual Localization Benchmark: <https://www.visuallocalization.net/>

²CVPR’20 Localization Challenge: <https://sites.google.com/view/vislocslamcvpr2020/localization-challenges>

³Winner of the CVPR’20 Localization Challenge: <https://github.com/cvg/Hierarchical-Localization>

While the Visual Localization Benchmark is well suited for visual localization in changing environments, it does not consider aspects due to the chosen datasets or metric that are important for visual place recognition. For example, the localization benchmark

- 1) does not investigate the influence of condition change without viewpoint changes and vice versa.
- 2) does not consider how condition changes *within* database or query affect methods.
- 3) evaluates only the global position of each query image.

In contrast, in a SLAM setup, the goal of place recognition might be to find all matchings between query and database images. On the one hand, this allows that query images need not to have a counterpart in the database and evaluates the ability to discover new unseen places. On the other hand, some place recognition algorithms exploit the knowledge that a query image has only a single matching database image. By evaluating all similarities between a query image and all database images, such assumptions can be revealed. These aspects are important for partial overlap between database and query as well as for loops and stops within database or query.

2) *The VPRiCE Challenge 2015:* This challenge⁴ was designed for a CVPR’15 workshop to benchmark and compare several submissions of place recognition algorithms. The benchmark is a concatenation of sequences from the datasets Nordland, Mapillary and QUT campus and city into one database and query set.

The results of an algorithm had to be submitted as actual image matches (e.g., “query 10 belongs to database 100”); i.e. the loop closure detection but not the place recognition performance is measured. Further, the submitted file must contain only two columns with a constantly increasing query image number. That means, there must not be loops or stops in the database (that have an overlap with the query).

3) *All-Day Visual Place Recognition Benchmark [18]:* This benchmark⁵ was proposed during the CVPR’15 workshop on visual place recognition in changing environments. The used datasets were recorded with a car equipped with RGB and thermal cameras as well as GPS and IMU around the Korea Advanced Institute of Science and Technology (KAIST) campus. While the benchmark may be interesting if thermal camera data is required, it covers merely a range of illumination changes in a single environment and is therefore not well suited for a comprehensive evaluation of visual place recognition algorithms over a variety of conditions and environments.

⁴The VPRiCE Challenge 2015: <https://roboticvision.atlassian.net/wiki/spaces/PUB/pages/14188617/The+VPRiCE+Challenge+2015+Visual+Place+Recognition+in+Changing+Environments>

⁵All-Day Visual Place Recognition Benchmark: <https://soonminhwang.github.io/publication/2015cvprwall-day-place-recognition/>

4) *Segway DRIVE Benchmark [87]:* This benchmark⁶ consists of several datasets collected by a fleet of delivery robots, and is designed for benchmarking place recognition and SLAM algorithms in indoor environments. Therefore, it covers appearance changes due to dynamic objects like people, long-term structural changes (e.g., removed furnitures), and illumination changes. Accordingly, it is unsuited as a benchmark that covers many kinds of condition changes in different large-scale environments.

5) *A realistic benchmark for visual indoor place recognition [88]:* Similarly to the Segway DRIVE Benchmark, the datasets of this benchmark were recorded in an indoor environment with two mobile robots and a tripod-camera, and covers dynamic objects, long-term changes and different lighting conditions. Due to the single environments and the restricted number of conditions, this dataset is unsuited as a large-scale, multi-condition benchmark.

⁶Segway DRIVE Benchmark: <http://drive.segwayrobotics.com/>