Beyond Holistic Descriptors, Keypoints and Fixed Patches: Multi-scale Superpixel Grids for Place Recognition in Changing Environments

Peer Neubert and Peter Protzel

***Abstract*—Vision-based place recognition in environments subject to severe appearance changes due to day-night cycles, changing weather or seasons is a challenging task. Existing methods typically exploit image sequences, holistic descriptors and/or training data. Each of these approaches limits the practical applicability, e.g. to constant viewpoints for usage of holistic image descriptors. Recently, the combination of local region detectors and descriptors based on Convolutional Neural Networks showed to be a promising approach to overcome these limitations. However, established region detectors, for example keypoint detectors, showed severe problems to provide repetitive landmarks despite dramatically changed appearance of the environment. Thus, they are typically replaced by holistic image descriptors or fixedly arranged patches - both known to be sensitive towards viewpoint changes. In this paper, we present a novel local region detector, SP-Grid, that is particularly suited for the combination of severe appearance and viewpoint changes. It is based on multi-scale image oversegmentations and is designed to combine the advantages of keypoints and fixed image patches by starting from an initial grid-like arrangement and subsequently adapting to the image content. The grid-like arrangement showed to be beneficial in the presence of severe appearance changes and the adaptation to the image content increases the robustness towards viewpoint changes. The experimental evaluation will show the benefit compared to existing local region detectors and holistic image descriptors.**

***Index Terms*—Localization, Visual-Based Navigation**

I. INTRODUCTION

ROBOTS operating autonomously over the course of days, weeks, and months have to cope with significant changes in the appearance of an environment. A single place can look extremely different dependent on the current season, weather conditions or time of day. Since state of the art algorithms for autonomous navigation are often based on vision and rely on the system's capability to recognize known places, such changes in the appearance pose a severe challenge for any robotic system aiming at autonomous long-term operation.

Fig. 1 shows a coarse taxonomy of existing approaches to visual place recognition in changing environments and how the proposed multiscale superpixel grid, SP-Grid, is related to them. Holistic image matching approaches compute a single descriptor for the whole image and showed to be very useful

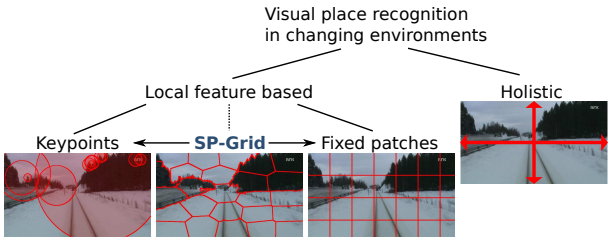

Fig. 1. A coarse taxonomy of approaches to visual place recognition in changing environments. The proposed multi-scale superpixel Grid, SP-Grid, is a local feature based approach combining the advantages of keypoints and fixed patches. Its regions cover the whole image and adapt to the image content.

in combination with image sequences [1] and for single image matching [2]. However, these holistic approaches are known to fail in the presence of viewpoint changes [3], [4].

Local feature based methods are known to be more robust to viewpoint changes [5]. The overview of place recognition approaches in Fig. 1 distinguishes in the field of *Local feature based* methods between *Keypoints* or *Fixed patches*.

Keypoints like SIFT, SURF or ORB are established components of successful localization systems, e.g. FAB-MAP [6]. Typically, a keypoint combines a local region detector and a descriptor. For example, SIFT uses a Difference-of-Gaussians approach to detect scale space extrema and gradient histograms for description. Appearance changes, as they happen, e.g., between day and nightfall, pose severe challenges for the *detection* and the *description* step of the keypoint features. Therefore, their application in the presence of environmental changes is known to be limited [7].

Recently, systems using descriptors based on Convolutional Neural Networks showed impressive performance for matching whole images [2] and local regions [4] despite severe appearance changes. Thus, they are a reasonable choice for a descriptor in changing environments - however, the question for suitable local region *detectors* remains open.

For place recognition in changing environments, patch- (or grid-) based methods showed impressive performance in the presence of severe appearance changes as they appear for example between "sunny summer days and stormy winter nights" [1], [8], [9]. The potential benefit is obvious: If no local region detector is involved in the place recognition, it cannot fail to detect corresponding regions.

However, the decoupling of the region detection from the image content by using a fixed grid of image patches, comes at the cost of reduced robustness to viewpoint changes. Dependent on the arrangement of the grid there are *critical*

Manuscript received: August 29, 2015; Revised November 20, 2015; Accepted December 12, 2015.
This paper was recommended for publication by Editor Cyrill Stachniss upon evaluation of the Associate Editor and Reviewers' comments.
The authors are with Faculty of Electrical Engineering and Information Technology, Technische Universität Chemnitz, Germany firstname.lastname@etit.tu-chemnitz.de
Digital Object Identifier (DOI): see top of this page.

cases, e.g. if the image content shifts horizontally for half the distance of two neighbouring patch centres. In this case, both neighbouring patches are maximally different from the new patch position.

In this paper, we propose a novel local region detector, the multiscale superpixel grid, SP-Grid. It is located between keypoints and fixed patches and tries to mitigate the disadvantages of both: keypoint detectors fail at detecting repetitive landmarks in the presence of severe appearance changes and fixed patches are sensitive to viewpoint changes. The proposed SP-Grid starts from an initial grid arrangement of image patches and adapts them to the image content using image oversegmentations, i.e. compact superpixel segmentations at multiple scales. The experimental results show that the proposed SP-Grid is more robust to viewpoint changes than a fixed grid and improves place recognition performance in case of severe appearance changes compared to available local image feature detectors (e.g. the scale space extrema used in SIFT). An open source implementation of the proposed approach will be available from our website[1].

II. RELATED WORK

The number of approaches to place recognition in changing environments grows rapidly. So far, no congruent solution for the practically relevant combination of severe appearance changes and changing viewpoints has been presented.

In terms of holistic approaches, SeqSLAM [1] combines sequence matching with a light weight image comparison front-end that builds on heavy image normalization and sums of absolute differences on a down sampled image (e.g. 64×32 pixels). These two components, using sequence and the image normalization, have also been used in other work. Continuous Appearance-based Trajectory SLAM (CAT-SLAM) [10] uses a particle filter with particle weighting based on local appearance and odometric similarity. Lowry et al. [11] use a combination of the underlying CAT-graph and a probabilistic whole image matching framework for place recognition in changing environments.

Badino et al. [12] implement the idea of visual sequence matching using a single SURF descriptor per image (WI-SURF) and Bayesian filtering on a topometric map. They show real-time localization on several 8 km tracks recorded at different seasons, times of day and illumination conditions.

Sequence Matching Across Route Traversals (SMART) [13] is another approach to extend the robustness of holistic image comparison based sequence matching towards varying view points and differences of the speed along the camera trajectories. It comprises a variable offset image matching to increase robustness against viewpoint changes and sample images at constant trajectory intervals, in contrast to constant time intervals, to handle varying speed between the two traversals of the same route. Therefore, a source of translational velocity is necessary. In [13], they used wheel encoders of the cars on which the cameras were mounted.

Johns and Yang [14] propose to quantise *local* features in both feature and image space to obtain discriminative statistics on the co-occurrences of features at different times of the day. They combine their approach with a sequence matching that can also handle non-zero acceleration and use local features to improve the robustness towards viewpoint changes. However, established solutions for non-changing environments based on local feature detection and local descriptors (e.g. the SURF keypoints of FAB-MAP) are known to reveal severe problems in changing environments [7], [13], [15].

An existing approach to overcome the problems of detecting repetitive features despite severe appearance changes is the usage of fixed patches. Naseer et al. [9] use a graph theoretical approach and formulate image matching as minimum cost flow problem in a data association graph. For computation of image similarities, they use a dense, regular grid of HOG descriptors and generate multiple route matching hypotheses by optimizing network flows. They show competitive results to SeqSLAM. In [8], Milford et al. combine a candidate selection based on whole image matching with a patch verification step based on local image regions. Since the candidates are selected based on the holistic approach, the local region matching can only mitigate the amount of false positive matchings, but not increase the number of matchings in the presence of viewpoint changes.

Another approach to dealing with severe appearance changes are learning-based methods, e.g. [15], [14], [16], [11]. While the idea to reason about environmental changes is appealing, the requirements on the necessary training data and/or knowledge about the environmental changes limit the generalisation capabilities and thus the practical applicability.

McManus et al. [17] propose scene signatures for localization. They represent each place in the database by a set of SVM classifiers that were trained to identify locally distinct rectangular image patches that are stable across different environmental conditions. In combination with a stereo camera they can estimate the relative pose between query and database images. Their system requires multiple images (e.g. 31) of the places that should be recognized and an initial coarse localization to select the set of scene signatures for which is searched in the query image.

Recently, holistic descriptors based on Convolutional Neural Networks (CNN) have been used for place recognition in changing environments [2], [18]. Sünderhauf et al. obtained image descriptors from the stacked output of a single CNN layer in [2]. They evaluated different layers and found the first convolutional layers to be the most robust against image changes, but sensitive to viewpoint changes. These descriptors showed impressive performance on a set of challenging datasets, including the cross-seasonal Nordland dataset.

Very recently, the CNN descriptors have been combined with local region detectors to increase the robustness towards viewpoint changes. In [3], an object proposal method, Edge-Boxes [19], is used to obtain the local regions. In our previous work [4], we combined several local region detectors with CNN-based descriptors, including the scale space extrema of SIFT [5], two object proposal methods [20], [21] and a segment soup [22]. The presented preliminary results indicate that the combination of local region detectors and CNN-based descriptors is promising - however, there is plenty of space

[1] https://www.tu-chemnitz.de/etit/proaut/forschung/cv/landmarks.html.en

for improvements on the repeated detection of local image regions in the presence of severe environmental changes. This work is continued in this paper by presenting a local region detector that outperforms the existing detectors in changing environments.

III. A NOVEL LOCAL REGION DETECTOR: SP-GRID

A. The underlying superpixel segmentations

The SP-Grid uses compact superpixel segmentations to create a set of overlapping regions at multiple scales. A superpixel segmentation is an oversegmentation of an image - or seen the other way around, a perceptual grouping of pixels. There exist various approaches to create superpixels with very different properties, please refer to [23], [24] for comparisons. Particularly *compact* superpixel segmentations create grid-like segments of regular sizes, shapes and distribution in the image. They are similar to fixed patches, but better aligned to the image content - and can thus increase the robustness towards viewpoint changes.

The SP-Grid region detector requires a compact superpixel algorithm that creates superpixels in a grid like arrangement, in particular with a defined 4-neighbourhood. Some of the available compact superpixel algorithms provide uniformly shaped segments but lack this neighbourhood. For example, SLIC [25] and Compact Watershed [26] are good superpixel choices for a SP-Grid. In the experiments presented here we use SLIC since it showed to provide slightly better segments than Compact Watershed. However, Compact Watershed is about twice as fast ($\sim$100 Hz @ (481$\times$321) pixels, desktop CPU) as the fastest available SLIC implementation [26].

B. Generating overlapping regions at multiple scales

For a non-overlapping grid layout, a compact superpixel segmentation can be used directly. To generate $(k \times k)$ regions, the image can be segmented into $(k \times k)$ superpixels and each superpixel becomes an output region. To allow for overlapping regions, a higher resolution of superpixels is computed and superpixels are subsequently grouped into regions. This is illustrated in Fig. 2. Starting from a (4×4) segmentation, all (2×2) groups of neighbouring superpixels are combined to obtain (3×3) regions.

To create SP-Grid regions at multiple scales, we compute an individual superpixel segmentation for each scale. It would also be possible to create regions at different scales from a single fine grained superpixel segmentation and group different numbers of superpixels (e.g. (2×2), (3×3) and so on) subsequently. Fig. 3 shows example images together with the (3×3), (4×4), (5×5), and (6×6) segmentations used to create the overlapping regions as is illustrated in Fig. 2.

C. The algorithmic implementation

Although the algorithmic concept is quite intuitive, details on the necessary algorithmic steps for extracting a set of SP-Grid regions from an image are given in Algorithm 1. The first step in line 1 is an initial rescaling of the image. Preliminary results (that are not shown here) indicate that to create segmentations of as few as (6×6) superpixels, reducing the image resolution by factor two is a reasonable choice for the datasets used here. The main loop in line 3 iterates over all scales of the grid. For each scale, the required number of superpixels is computed from the width and height of a region and the overlap of regions, both measured in number of superpixels (lines 5 and 6). For the example regions in Fig. 2, the width and height of a region is 2 superpixels and the overlap is 1 superpixel. Changing these numbers can be used to vary the amount of overlap and to generate regions of different sizes from a single superpixel segmentation.

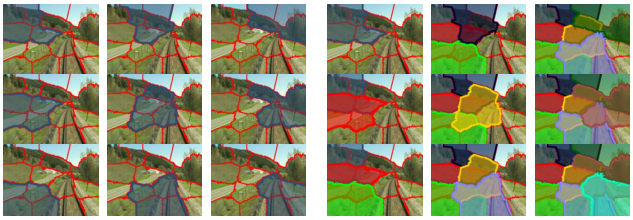

Fig. 2. *(left)* Illustration of the superpixel groupings to obtain the SP-Grid regions for the (3×3) layer. All (2×2) groups of neighbouring superpixels are combined and result in 9 regions. *(right)* Visualization of the overlap.

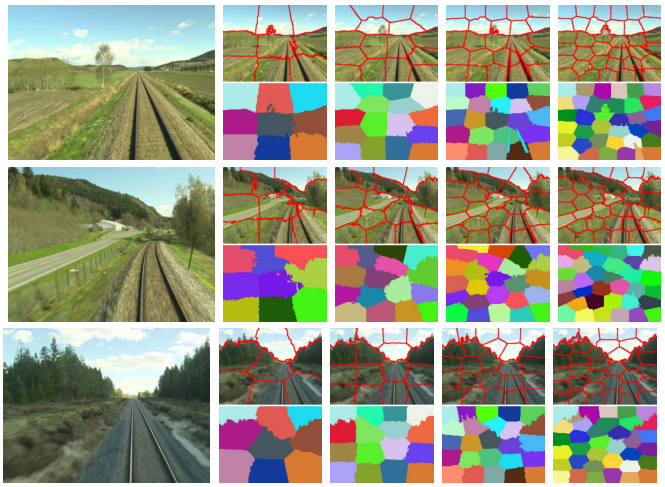

Fig. 3. Example Nordland images and resulting SP-Grid superpixel layers.

In line 7 the superpixel segmentation is computed. The remaining lines 8-20 of the main loop are dedicated to the computation of the regions using the superpixel label image. For the conducted index arithmetic, it is assumed that the superpixels are arranged grid-like and the labels are in column major order. For each region (x, y), the superpixel label index at the top left corner is computed (lines 10 and 11) and subsequently all $(nSpPerRegion \times nSpPerRegion)$ labels are collected (lines 12-17). Line 18 collects the resulting regions. They can be obtained from a merging of the assigned superpixel areas or by any other combination (e.g. simply the bounding box containing all assigned superpixels).

This algorithmic description can create different arrangements of regions at different scales and amounts of overlap. While an extensive evaluation of all degrees of freedom is beyond the scope of this paper, section V-A will provide a set of experiments to find a *reasonable* configuration.

IV. EXPERIMENTAL SETUP

A. The compared approaches

In section V-C we will compare the novel SP-Grid with other types of local region detectors that showed promising

Algorithm 1: Outline of the SP-Grid algorithm.

Data: Image
Grid resolution for each scale: Grid
Width and height of a region in superpixels: nSpPerRegion
Overlap of regions in superpixels: spOverlap
Result: Set of regions: R

```
(Optionally) Resize the image;
Initialize empty set of regions R=∅;
foreach Scale level s of the grid configuration do
    // Get grid resolution for the current scale level
    (nx, ny) = getGridSize(Grid, s);
    // Get size of the superpixel segmentation
    nSpX = nx · nSpPerRegion - (nx-1) · spOverlap;
    nSpY = ny · nSpPerRegion - (ny-1) · spOverlap;
    // Compute the superpixel label image, superpixel labels
    // have to be arranged column major order
    L = performSuperpixelSegmentation(I, nSpX, nSpY);
    // Create regions by collecting (nSpPerRegion × nSpPerRegion)
    // groups of segments
    for y=1:ny do
        for x=1:nx do
            // Get superpixel label coordinates of top left corner
            xTL = (x-1) · (nSPperRegion-spOverlap)+1;
            yTL = (y-1) · (nSPperRegion-spOverlap))+1;
            // Collect all superpixel labels for this region
            labels = ∅;
            for yy=yTL:yTL+nSpPerRegion-1 do
                for xx=xTL:xTL+nSpPerRegion-1 do
                    labels = labels ∪ {(xx-1) · nSpY + yy};
                end
            end
            // Compute region from superpixel labels and include in R
            R = R ∪ {getRegionFromSuperpixels(L,labels)};
        end
    end
end
```

results in [4] and [3]:

- The scale space extrema detector based on differences of Gaussians as it is used for SIFT (with the typical scale factor six), named SIFT-DoG-6 in the experiments.
- A segment soup SP-Soup [4], based on multiple segmentations using the segmentation algorithm from [27].
- Object proposal algorithms: EdgeBoxes [19], was used in combination with CNN descriptors in [3]. Multiscale Combinatorial Grouping (MCG) [28] is another recently presented, promising object proposal algorithm.
- Finally, we also include the holistic CNN descriptor that is computed according to [2] but using the same CNN network as for the local regions.

B. The image matching procedure: CNN and Star-Hough

To apply the SP-Grid and the other local region descriptors for place recognition, they have to be combined with a descriptor and an image matching scheme that computes image similarities from the local regions and their descriptors. We use the same methodology introduced in [4]: A conv3-layer descriptor is computed for the bounding box of each local region using the VGG-M network.[2] The descriptors are compared using the cosine distance metric. The pairwise region comparisons of the landmark sets from two images are combined using the Star-Hough image matching procedure. Star-Hough incorporates the spatial arrangement of the landmarks in the image by evaluating votes for a shift of the centre of the Star Graph Model created by the landmarks in each image. It accomplishes a similar task like outlier rejection based on epipolar geometry, but is particularly suited for landmarks with low precision of their spatial position and high rates of outlier matchings - both issues typically occur in changing environments and both are known to challenge epipolar geometry estimation [5]. Example SP-Grid region matchings can be seen in Fig. 4. Please refer to [4] for details on the image matching procedure.

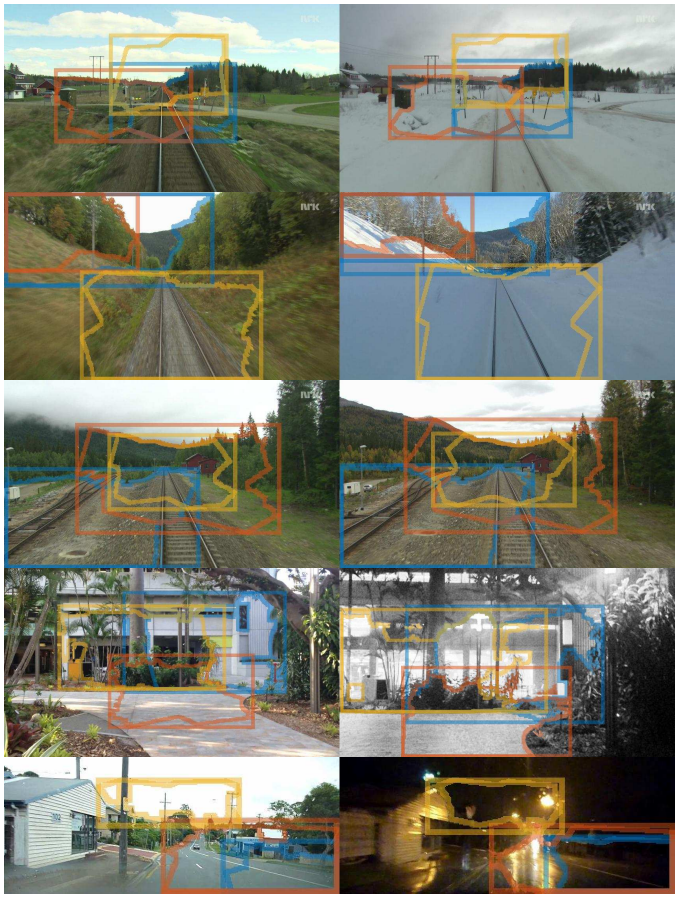

Fig. 4. Example matchings of SP-Grid regions. Each row shows two images of the same place from Nordland Spring-Winter, Fall-Winter, Summer-Fall, GardensPoint dayLeft-nightRight and Alderley datsets. The three best matching regions are visualized with the same colour, the bounding boxes show the image patch that is described by the the CNN.

C. The evaluation methodology: Precision-recall curves

We present experimental results on place recognition using precision recall curves. While this is an often used performance measure, details on the particular evaluation methodology have a large influence on the resulting curves. This has to be considered when comparing results from different papers as will be demonstrated in Fig. 9 and 11 and discuss in section VI.

[2]We use the implementation from http://www.vlfeat.org/

TABLE I
DEFAULT PARAMETERS FOR THE SP-GRID

Parameter	Value
Multi-scale grid resolutions	(1×1), (2×2), (3×3), (4×4), (5×5)
Resulting number of regions	55
Superpixel algorithm	SLIC implementation from VLFeat (compactness=30k)
Width and height of a region	2 superpixels
Region overlap	1 superpixel
Image rescaling factor	0.5

TABLE II
EXAMPLE CONFIGURATIONS TO OBTAIN A CERTAIN OVERLAP.

nSPperRegion	spOverlap	Resulting Overlap
2	0	no overlap
2	1	50 %
10	9	90 %

Given the image similarities between all possible image pairings, they are divided into matchings and non-matchings by applying a threshold t on this similarity. All image matchings that correspond to a ground truth place correspondence are counted as true positives, all matchings that do not show the same place according to the ground truth are considered false positives, and false negatives are all image pairings of the ground truth that are not in the set of matchings. From these three values, a point on the precision-recall curve is computed. To obtain the curve, the threshold t is varied.

D. The datasets: Nordland, GardensPoint and Alderley

The experiments are conducted using three datasets: The Nordland dataset comprises images of all *seasons* from four journeys on a 728 km train route across Norway [15]. We evaluate on the complete journey of the test dataset [15] and a unique place is assumed each 10 frames. The images are synchronized and aligned: the pixel (i,j) of the n-th frame of the summer sequence approximately corresponds to the pixel (i,j) in the n-th frame of the winter sequence. For parameter evaluation, a smaller set of 186 uniformly sampled places from the Spring-Winter validation dataset [15] is used.

The GardensPoint dataset[3] provides images captured from a hand held camera at three traversals of an mixed indoor and outdoor route. There are two traversals at *day* and one at *night*. The first daytime traversal and the night traversal are on the right side of the path, the second daytime run is on the left side of the path. The most challenging dataset is the Alderley dataset [1], comprising images from a summer day and a rainy night captured from a driving car. We took every 100th frame of this dataset (on average that corresponds to a few ten meters). Example images of all datasets will be shown together with the resulting curves of the place recognition experiments.

The datasets provide different types of environmental changes and they are also quite different in terms of the amount of contained viewpoint change. The Nordland images are pixel aligned. They will be used directly and with additional synthetic lateral image shifts. The effect of a 10% shift can roughly be compared to the rotation of the used camera of about 5 degrees. The lateral distance between the camera trajectories on the GardensPoint dataset depends on the actual width of the path and varies between 1-4 meters. The Alderley images provide real world viewpoint changes as they happen during two traversals of the same route with a car in real traffic.

[3]Recorded by Arren Glover, https://wiki.qut.edu.au/display/cyphy/Day+and+Night+with+Lateral+Pose+Change+Datasets

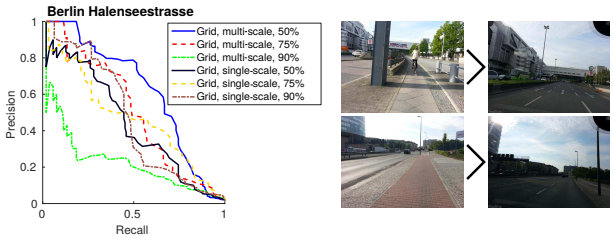

Fig. 5. Parameters I: Evaluation of the grid arrangement on the Berlin Halenseestrasse dataset and example images of two places. Each setup provides about 50 regions using a single or multiple scales and with different overlap between neighbouring regions (given in % of the region width).

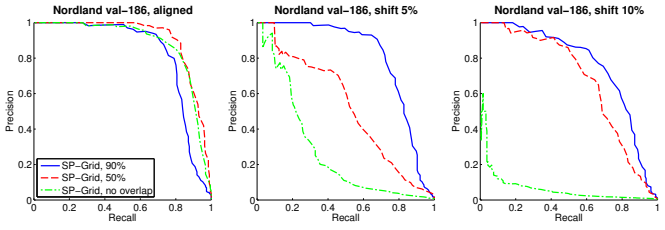

Fig. 6. Parameters II: Influence of SP-Grid region overlap on Nordland Sping-Winter with 0, 5% and 10% horizontal image shift. 50% and 90% region overlap are reasonable choices.

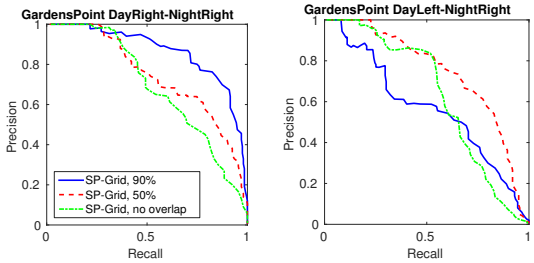

Fig. 7. Parameters III: Same as Fig. 6 but on different variants of the GardensPoint dataset. In case of the lateral shift at DayLeft-NightRight, the overlap of 50% is preferable.

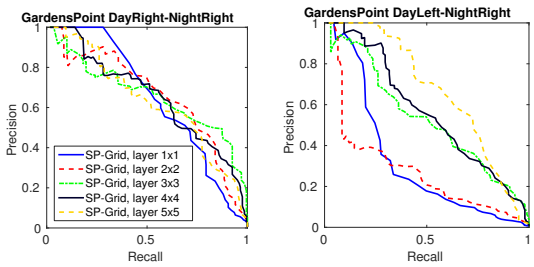

Fig. 8. Parameters IV: Contribution of different SP-Grid region levels. For (roughly) aligned images (left), all regions at all scale levels perform similar. In case of additional viewpoint change (right), smaller scales perform better.

V. EXPERIMENTAL RESULTS

A. Parameter selection

While an extensive evaluation of all parameters is beyond the scope of this paper, we want to find a reasonable configuration that can be used for comparing SP-Grid to existing approaches. Besides the choice of the superpixel algorithm (cf. section III-A), the arrangement of the SP-Grid regions is

important. It results from three decisions: the used scales, the number of regions per scale and their overlap.

To evaluate the arrangement independent from the superpixel algorithm and the subsequent comparison to existing approaches, we conduct an initial set of experiments using a fixed grid (Grid) on the Berlin Halenseestrasse dataset [3]. The fixed grid is obtained by replacing the superpixels with static rectangular patches in algorithm 1. At the Berlin Halenseestrasse dataset, the viewpoint changes between the camera mounted behind the windscreen of a car driving on a street and a camera on a bicycle on the cycle lane alongside the road. Additionally, the illumination changes due to different position of the sun (cf. Fig. 5).

In accordance with the experiments using CNN-based landmarks in [4] and [3], the total number of local regions should be about 50. Fig. 5 shows evaluation results using single-scale (7×7 regions) and multi-scale ($(5 \times 5) + (4 \times 4) + (3 \times 3) + (2 \times 2) + (1 \times 1)$ regions) setups with different amounts of overlap given in % of the region width. Since the used CNN descriptors are not scale invariant, using regions at multiple scales performs better for these severe viewpoint changes including scale changes. The combination with 50% overlap between regions provides the best results.

Table I lists the default parameters that will be used in the following. An extensive evaluation of these degrees of freedom is beyond the scope of this paper, however, we want to present some insights on the most important parameters.

For the SP-Grid as described in algorithm 1, the number of regions on a layer is given by (*nx, ny*) and the overlap is controlled by the number of superpixels per regions (*nSPperRegion*) and the number of common superpixels between neighbouring regions (*spOverlap*). Different settings of these parameters and the resulting overlaps are listed in table II. Fig. 6 and 7 show place recognition results using these settings for the Nordland validation dataset with different amounts of synthetic viewpoint change and the GardensPoint dataset (with and without lateral shift).

It can be seen that overlapping regions are preferable and 50% and 90% overlap are reasonable choices. While 90%-overlapping regions show the better performance on the synthetic viewpoint changes of the Nordland dataset, the 50%-overlapping regions are better at the real world viewpoint changes of the Halenseestrasse (Fig. 5) and GardensPoint (Fig. 7) datasets.

Fig. 8 compares the contribution of the regions at different scale levels. For (roughly) aligned images the performance of all scale levels is comparable. In case of additional viewpoint changes the larger (1×1) and (2×2) regions are more affected by the changing image boundaries and thus perform worse.

Dependent on the knowledge about motion constraints or properties of the environment, a different parameter set may be preferable. However, the setup shown in table I is a *reasonable* configuration that can deal with the viewpoint changes that typically occur in the available datasets.

B. Is the superpixel grid better than a fixed grid?

From a theoretical point of view, the SP-Grid regions are better aligned to the image content and should thus be more robust to viewpoint changes than the patches of a fixedly arranged grid. To evaluate whether there is a real existing benefit when using the SP-Grid instead of the fixed Grid, Fig. 9 shows the results of both algorithms on place recognition on the GardensPoint datasets. An example image triple showing the appearance change and the lateral shift can be seen on the right part of Fig. 9. The resulting curves for the DayLeft-NightRight comparison show an improvement in F-score and recall at 100% precision when using the SP-Grid. The overall place recognition performance (of course) increases with less restrictive distances. SP-Grid always outperforms Grid in the presence of viewpoint changes. The right plot shows the results on the GardensPoint dataset without viewpoint change - there, the fixedly arranged Grid performs better.

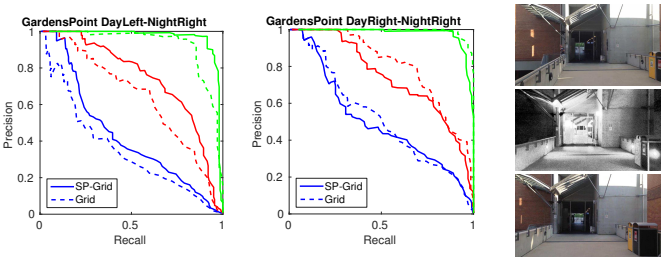

Fig. 9. Comparison of Grid (dashed) and SP-Grid (solid) on the Gardens Point datasets (with and without lateral shift). The different colours indicate how restrictively image matchings are accepted, the maximum image distance is: blue≙1, red≙3, and green≙10. Please see section VI for details.

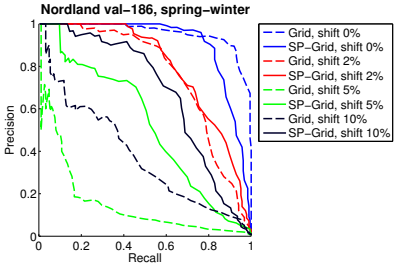

Fig. 10. Comparison of the robustness of Grid and SP-Grid towards different amounts of horizontal shifts in corresponding images. For shifts $> 0\%$ SP-Grid performs better than Grid in this place recognition experiment. A shift of 5% is the most critical case from this comparison.

To more precisely evaluate the influence of viewpoint changes, we can use the aligned Nordland validation dataset in combination with artificial viewpoint changes (similar to Fig. 6). The place recognition results for different amounts of horizontal shift can be seen in Fig. 10. The SP-Grid clearly outperforms the fixed Grid for all non zero amounts of shift in terms of F-score and recall at 100% precision. For the chosen (and presumably not optimal) choice of the initial grid arrangement described in section V-A, a shift of 5% of the image width constitutes a *critical* case. The regions of the fixedly arranged Grid show only small overlap for this shift. For larger shifts, the overlap increases since regions now overlap with the regions corresponding to their neighbours in the grid. This is a periodic behaviour with different frequency for each layer of the grid. The SP-Grid smooths the average overlap compared to the fixed Grid since the superpixels adapt the initial grid to the image content. While this may decrease the performance for perfectly aligned images, place recognition based on the SP-Grid shows to be more robust against viewpoint changes - in particular for the critical shifts.

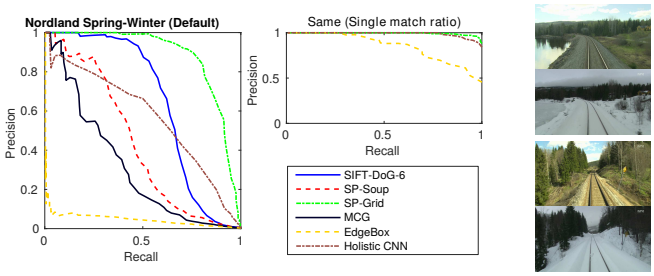

Fig. 11. Results on Nordland Spring-Winter. To allow a comparison with results from the literature (in particular [3]), the plot in the middle shows the *same results* with a different method to compute the precision recall curves, see section VI for details.

C. Comparison of SP-Grid to other local region detectors

We compare the place recognition performance of SP-Grid with the existing approaches presented in sec. IV. Fig. 11-12 show results on different seasonal combinations on the full journey of the Nordland test dataset. The SP-Grid provides higher F-score and better recall at 100% precision than the compared approaches. According to [29], fall-winter is the hardest and summer-fall the easiest seasonal combination. The benefit for using SP-Grid increases with increasing severity of the appearance change due to seasonal change.

Fig. 13 demonstrates the influence of an artificial viewpoint change by shifting the images 5% of the image width - this showed to be the most critical case for SP-Grid in the previous experiments shown in Fig. 10. The performance of SP-Grid clearly drops but is significantly more stable than the holistic approach even in this particular challenging configuration for SP-Grid. The performance of the other region detectors drops slightly while keeping the order from the aligned setup. SIFT-DoG-6 is almost not affected by this shift and performs better than SP-Grid in this setup.

Fig. 14 shows results on the GardensPoint dataset. The appearance changes from day to night challenge all region detectors. SP-Grid provides the best results in the setup including the lateral shift but is outperformed by EdgeBoxes on the roughly aligned dataset. The results of both object-proposal algorithms (EdgeBoxes and MCG) vary between both setups of this dataset. They are intended to find regions that are likely to contain an object in an image, they are not designed to find repetitive landmarks. The example image triple shown on the right side of this figure shows that due to the viewpoint change, salient objects like the bench and the table on the very right can disappear. This might also remove false positive matchings of MCG and thus cause an improved precision. Moreover, the route of the GardensPoint dataset is rather small compared to the other datasets, which makes it sensitive to individual effects on few images.

A significantly larger track is provided by the Alderley dataset. This dataset comprises two drives through a suburb, one at a sunny day and the other during a rainy night with low visibility. Example images of these severe appearance changes can be seen in the right part of Fig. 15. This is the most challenging dataset in this evaluation. The benefit from using the SP-Grid regions further increases, but the absolute performance in terms of F-score and recall at 100% precision is worse than for the other datasets.

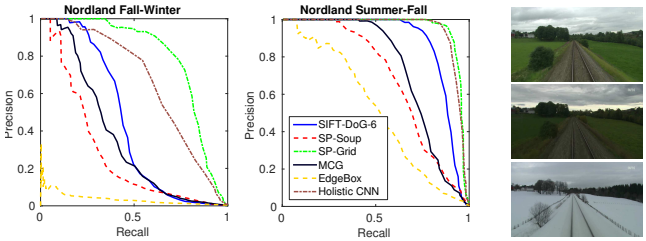

Fig. 12. Results on the Nordland fall-winter and summer-fall combinations. Example images top to bottom: summer, fall, winter.

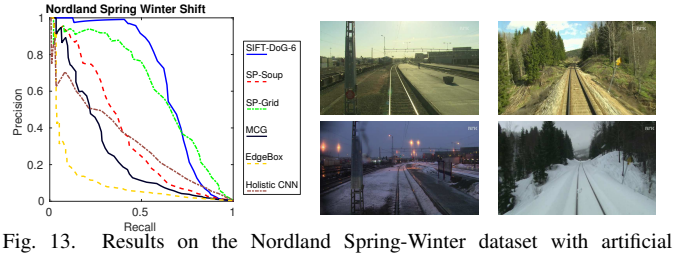

Fig. 13. Results on the Nordland Spring-Winter dataset with artificial viewpoint changes. The images on the right show this rather small shift of 5% of the image width.

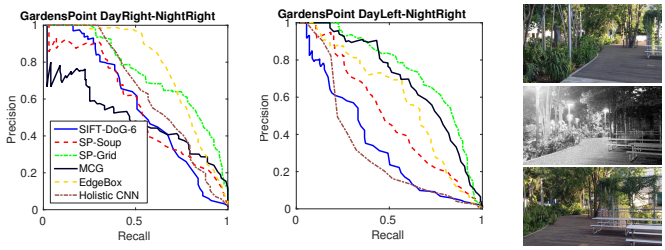

Fig. 14. Results on GardensPoint day-night datasets (max. distance = 3).

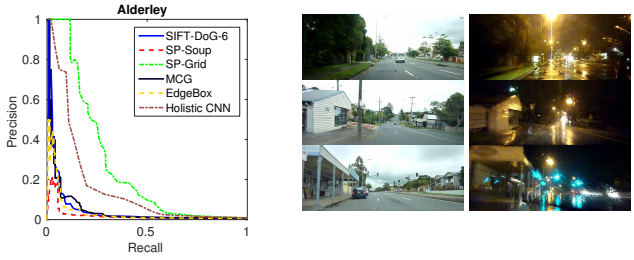

Fig. 15. Results on Alderley - the most challenging dataset.

For these low visibility conditions, the SP-Grid regions converge to a regular grid in the absence of considerable image gradients due to the compactness constraints of the used superpixel algorithm - this constitutes a reasonable default set of regions. In particular, this is in contrast to the SP-Soup that adapts to even very small image gradients due to the absence of compactness constraints.

VI. Two Notes on the Comparability of Place Recognition Results

The different colours in Fig. 9 indicate the maximum distance of matched frames in the sequence that are accepted as showing the same place. GardensPoint and the other used datasets comprise image sequences, not images of disjunct places. Typically this is handled by subsampling a set of disjunct places or by allowing matching images up to a certain distance in the original sequence - the effect of this distance can be seen by the differently coloured curves in Fig. 9. Such variations in the evaluation have to be considered when comparing results from different papers.

A second potential problem is illustrated in Fig. 11: The plot in the middle shows a second set of precision recall curves using the *same* image similarities as the left plot. However, these curves indicate a significantly better performance of EdgeBoxes (while the proposed SP-Grid still performs significantly better). This second plot is generated with the evaluation method used in [3] and similar to the there presented results of EdgeBoxes and CNN on this dataset. The difference between the two evaluation methods is not in the computation of image similarities - but how true and false matchings obtained from the image similarities are computed. In [3], only a *single* matching for each query image is allowed and the decision threshold (to classify matchings and non-matchings) is not applied on the image similarity but on the ratio of the similarities of the best to the second best matching. While neither of the two evaluation methods can be considered generally better or worse (however, our method allows multiple revisits of a place and the method from [3] does not), they produce fundamentally different curves. Thus, the results can not be compared between different papers directly.

VII. CONCLUSIONS

This paper started from a discussion of the need for place recognition approaches that can deal with severe appearance *and* viewpoint changes. The combination of local region detectors and CNN-based descriptors showed promising results in previous work by us and others. However, in the presence of severe appearance changes (e.g. between day and night) existing local region *detectors* reveal severe problems. We proposed a novel region detector, SP-Grid, that is located somewhere between keypoints and fixedly arranged patches and is designed to combine advantages of both. The experimental evaluation showed that the SP-Grid can mitigate the negative influence of critical viewpoint changes compared to fixedly arranged patches and provides considerably better performance than the compared methods (e.g. SIFT-DoG-6) in case of severe appearance changes. The presented implementation of the concept of a SP-Grid includes some components and parameters for which a reasonable setup was chosen, but presumably not an optimal. Particularly an adaptation of the initial grid arrangement based on knowledge about the expected viewpoint changes and a superpixel algorithm, that is particularly designed for this system, may further improve the results.

REFERENCES

[1] M. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights." in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2012.

[2] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," *Computing Research Repository (CoRR)*, vol. abs/1501.04158, 2015. [Online]. Available: http://arxiv.org/abs/1501.04158

[3] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. of Robotics: Science and Systems*, Rome, Italy, July 2015.

[4] P. Neubert and P. Protzel, "Local region detector + CNN based landmarks for practical place recognition in changing environments," in *Proc. of European Conference on Mobile Robotics (ECMR)*, 2015.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, 2004.

[6] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[7] C. Valgren and A. J. Lilienthal, "Sift, surf & seasons: Appearance-based long-term localization in outdoor environments," *Robotics and Autonomous Systems*, vol. 58, no. 2, pp. 149–156, 2010.

[8] M. Milford, E. Vig, W. Scheirer, and D. Cox, "Towards condition-invariant, top-down visual place recognition," in *Proc. of Australasian Conference on Robotics and Automation (ACRA)*, 2013.

[9] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows." in *Proc. of AAAI Conference on Artificial Intelligence*, 2014.

[10] W. Maddern and S. Vidas, "Towards robust night and day place recognition using visible and thermal imaging," in *Proc. of Robotics Science and Systems Conference (RSS)*, 2012.

[11] S. M. Lowry, G. F. Wyeth, and M. Milford, "Towards training-free appearance-based localization: Probabilistic models for whole-image descriptors," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[12] H. Badino, D. F. Huber, and T. Kanade, "Real-time topometric localization." in *Proc. of International Conference on Robotics and Automation (ICRA)*, 2012.

[13] E. Pepperell, P. Corke, and M. Milford, "All-environment visual place recognition with smart," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 1612–1618.

[14] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day." in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

[15] P. Neubert, N. Sünderhauf, and P. Protzel, "Superpixel-based appearance change prediction for long-term navigation across seasons," *Journal of Robotics and Autonomous Systems*, vol. 69, no. 0, pp. 15 – 27, 2015.

[16] A. Ranganathan, S. Matsumoto, and D. Ilstrup, "Towards illumination invariance for visual localization," in *In proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2013, pp. 3791–3798.

[17] C. McManus, B. Upcroft, and P. Newman, "Scene signatures: Localised and point-less features for localisation," in *Proceedings of Robotics Science and Systems (RSS)*, Berkley, CA, USA, July 2014.

[18] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," in *In proceedings of the Australasian Conference on Robotics and Automation*, December 2014. [Online]. Available: http://eprints.qut.edu.au/79662/

[19] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision (ECCV)*. Proc. of European Conference on Computer Vision (ECCV), September 2014.

[20] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 11, 2012.

[21] S. Manen, M. Guillaumin, and L. Van Gool, "Prime object proposals with randomized prim's algorithm," in *Proc. of International Conference on Computer Vision (ICCV)*, 2013.

[22] T. Malisiewicz and A. A. Efros, "Improving spatial support for objects via multiple segmentations," in *Proc. of British Machine Vision Conference (BMVC)*, 2007.

[23] P. Neubert and P. Protzel, "Evaluating superpixels in video: Metrics beyond figure-ground segmentation," in *Proc. of British Machine Vision Conference (BMVC)*, 2013.

[24] ——, "Superpixel benchmark and comparison," in *Proc. of Forum Bildverarbeitung*, 2012.

[25] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 11, pp. 2274–2282, 2012.

[26] P. Neubert and P. Protzel, "Compact Watershed and Preemptive SLIC: On improving trade-offs of superpixel segmentation algorithms," in *Proc. of International Conference on Pattern Recognition (ICPR)*, 2014.

[27] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, 2004.

[28] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marqués, and J. Malik, "Multiscale combinatorial grouping," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[29] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons." in *Proc. of IEEE International Conference on Robotics and Automation (ICRA), Workshop on Long-Term Autonomy*, 2013.