Local Region Detector + CNN based Landmarks for Practical Place Recognition in Changing Environments

Peer Neubert and Peter Protzel
Technische Universität Chemnitz
09126 Chemnitz, Germany
Contact: peer.neubert@etit.tu-chemnitz.de

***Abstract*—Visual place recognition is a mature field in mobile robotics research. Recognizing places in datasets covering traversals of hundreds or thousands of kilometres and accurate localization in small and medium size environments have been successfully demonstrated. However, for real world long term operation, visual place recognition has to face severe environmental appearance changes due to day-night cycles, seasonal or weather changes. Existing approaches for recognizing places in such changing environments provide solutions for matching images from the exact same viewpoint using powerful holistic descriptors, using less sophisticated holistic descriptors in combinations with images sequences, and/or pose strong requirements on training data to learn systematic appearance changes. In this paper, we present a novel, training free, single image matching procedure that builds upon local region detectors for powerful Convolutional Neural Network (CNN) based descriptors. It can be used with a broad range of local region detectors including keypoints, segmentation based approaches and object proposals. We propose a novel evaluation criterion for selection of an appropriate local region detector for changing environments and compare several available detectors. The scale space extrema detector known from the SIFT keypoint detector in combination with appropriate magnification factors performs best. We present preliminary results of the proposed image matching procedure with several region detectors on the challenging Nordland dataset on place recognition between different seasons and a dataset including severe viewpoint changes. The proposed method outperforms the best existing holistic method for place recognition in such changing environments and can additionally handle severe viewpoint changes. Additionally, the combination of the best performing detectors with superpixel based spatial image support shows promising results.**

I. Introduction

Visual place recognition in changing environments describes the problem of matching images (or image sequences) of places that are subject to severe appearance changes induced by day-night cycles, weather or seasonal changes. Two images showing an example scene from the Nordland dataset can be seen in Fig. 1. This is a challenging and active research field for mobile robotics as well as computer vision. Visual localization has benefited a lot from the evolution of local keypoint detectors and descriptors like Harris corners or SIFT. For pure place recognition there are also holistic image descriptors like GIST. However, these established methods show problems in the presence of severe image changes, e.g. matching places seen in summer when revisiting in winter.

There have been different approaches for place recognition in changing environments, e.g. based on matching image sequences (e.g. SeqSLAM [1]), exploiting training data to predict systematic appearance changes [2], or using sophisticated holistic descriptors obtained from Convolutional Neural Networks (CNN) [3]. However, none of the existing approaches provides place recognition in changing environments in a practical setup including viewpoint changes. The authors of [4],[3] demonstrated these problems on datasets with shifted and cropped images.

978-1-4673-9163-4/15/$31.00 © 2015 IEEE

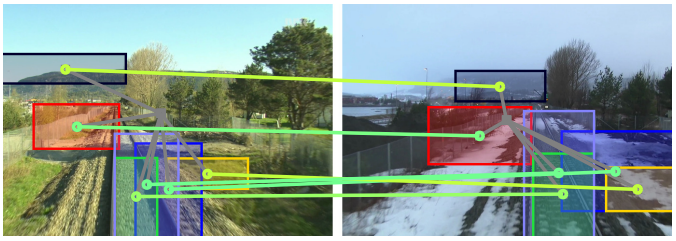

Fig. 1. Illustration of our novel, training-free approach for single image place recognition in changing environments (e.g. between summer and winter). It combines the robustness against appearance changes of CNN descriptors with the robustness to viewpoint changes of local landmarks. It can be used with a broad range of local region detectors including keypoints, segmentation based approaches and object proposals.

We present a novel approach to place recognition in changing environments in the presence of viewpoint changes based on local image features. Using local image features has proven to be a successful approach facing changing viewpoints. These local features are composed of a feature detection and a feature description step. The challenge for applications in changing environments is to find detectors and descriptors which both can handle severe appearance changes and can be combined to provide useful landmarks.

After a short overview of related work in the following section, we address this challenging place recognition task in two steps. In a first step, we discuss the requirements for the feature detector and present a measure and experiments to evaluate the repeatability of feature detectors in changing environments in Section III. Further, we compare a selection of different detectors including keypoint detectors, oversegmentation based approaches and object proposal methods.

Section IV discusses the second step: feature description. We use the output of the best performing detector as spatial image support for the CNN features that have shown to be robust to seasonal changes [3]. We present a training free, single image matching scheme using these landmarks based on simple star graph models [5]. Section V shows how the

proposed approach outperforms the state-of-the-art method on place recognition experiments on the challenging Nordland dataset and the Berlin dataset including severe viewpoint changes. Results and an open source implementation of the proposed approach are available from our website.[1]

II. Related Work

Traditionally, visual place recognition is either based on matching local features (like SIFT keypoints [6]), bags of visual words (like FAB-MAP [7]), holistic image descriptors (like GIST [8]), or combinations. In changing environments, one can either try to organize the different appearances, e.g. in form of a plastic map [9], or to recognize places despite these differences. While a complete review of existing approaches is beyond the scope of this paper, we want to present some related approaches.

In terms of local features, Valgren and Lilienthal [10] used SIFT and U-SURF keypoints for place recognition across seasons. They show high recognition rates on a small dataset with omnidirectional images of five places from different seasons. Milford et al. provided several approaches including SeqSLAM [1]. It builds upon a lightweight holistic visual matching front-end and explicitly matches local sequences of images. They show impressive results on matching challenging scenes across seasons, time of day and weather conditions. Since it builds upon a holistic image representation, SeqSLAM is sensitive to viewpoint changes [4]. In prior work [2], we used superpixel vocabularies to learn and predict systematic appearance changes. We showed how this prediction step can improve the performance of subsequent place recognition techniques like SeqSLAM. Since the predicted images are not suitable for local keypoint extraction due to smoothing and artefacts, the same issues of holistic image matching as for SeqSLAM holds for this approach.

In recent years, image descriptors based on Convolutional Neural Networks (CNN) showed impressive performance on a variety of computer vision tasks, most obviously object recognition [11]. Sünderhauf et al. [3] proposed to use descriptors obtained from CNNs for place recognition in changing environments. They obtained image descriptors from the stacked output of a single CNN layer. They evaluated different layers and found the lower convolutional layers to be the most robust against image changes, but sensitive to viewpoint changes. They also proposed to compute multiple descriptors using a sliding window scheme to improve the performance under changing viewpoints. However, the general drawbacks of sliding window approaches (high number of windows, sensitivity to partial occlusions, and dynamic objects) still apply. Nevertheless, these CNN descriptors showed impressive performance on a set of challenging datasets, including the cross-seasonal Nordland dataset.

In the following, we will investigate how these powerful CNN descriptors can be combined with local spatial image support to become more robust against viewpoint changes and provide practical landmarks for changing environments.

III. Detecting Repetitive Landmarks in Changing Environments

Landmarks in changing environments require repeated detection and sophisticated descriptors. In this section we will first provide a measure to evaluate the repeatability of region detectors in the presence of severe image changes (Sec. III-A), followed by the description of potential region detectors including keypoints, oversegmentations and object proposals (Sec. III-B), and finally an experimental comparison of these detectors (Sec. III-C). The best performing detectors are then used in the matching scheme presented in Sec. IV for the place recognition experiments in Sec. V.

A. How to Measure the Repeatability of Feature Detectors

In their influential work Mikolajczyk et al. [12] provide a methodology to evaluate the repeatability of feature detectors in terms of the *localization accuracy*. We build upon their methodology and adapt it for the requirements on landmarks under severe appearance changes to measure the amount of *common spatial image support*. In principle their approach is the following: Given two images I_A, I_B of the same scene and the transformation T_{AB} between these images (i.e. the ground truth optical flow), they detect features in both images and evaluate their overlap. Given the sets of pixels P_A, P_B constituting the spatial image support of two features f_A^i from I_A and f_B^j from I_B, the overlap is computed as the "intersection over union" (*IoU*):

$$IoU(f_A^i, f_B^j) = \frac{|T_{AB}(P_A^i) \cap P_B^j|}{|T_{AB}(P_A^i) \cup P_B^j|} \tag{1}$$

E.g. f_A^i, f_B^j may be SIFT features, P_A^i, P_B^j are all pixels belonging to the corresponding ellipses, and $T_{AB}(P_A^i)$ are the pixels of f_A^i moved to the image space of I_B.

An important step in their evaluation is a normalization with respect to the features size. They show that the performance of a region detector in their measure can be improved by simply increasing the size of the detected regions. Therefore, they compute for each feature comparison $IoU(f_A^i, f_B^j)$ a rescaling factor that normalizes the size of f_A^i to a given diameter and apply this rescaling on both features. This makes different region detectors comparable. However, the authors clearly point out that such a normalization should not be applied for real applications of these feature detectors.

Our experiments with image features for place recognition in changing environments support the hypotheses of a real dependency of the features' sizes and the resulting place recognition performance. E.g. typical seasonal changes induce severe appearance differences at small scales: leaves change their shape and colour, snow and ice cover meadows, rain modifies the reflective properties of surfaces and so on. In contrast, the appearance of coarser structures is supposed to be more persistent: mountains, buildings, streets and trees change their detailed appearance but are likely to remain at the same global configuration.

Therefore, we propose to use the same *IoU* criterion as in [12] but in combination with another normalization procedure to take the feature size into account. The objective is to measure an upper bound on feature matching performance that can be achieved with a given feature detector. Given two sets of detections $F_A = \{f_A^1, f_A^2, ...), F_B = \{f_B^1, f_B^2, ...)$, from two images with known optical flow T_{AB}, we transfer all features to the same image space and assign to each feature of F_A the maximum IoU between this feature and a feature from F_B:

$$IoU_{max}(f_A^i, F_B) = \max_{f_B^j \in F_B} IoU(f_A^i, f_B^j) \tag{2}$$

These pairs are the combinations with the highest rate on common pixels and thus (approximately) the highest rate on common world points. This is supposed to represent an

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

upper bound for what could be found by matching based on real image descriptors. If there are no feature pairs detected showing sufficient overlap, the best feature descriptor will fail to match the images.

As a measure for the repeatability, we can then simply count for a set of image pairs the average number of features whose IoU_{max} exceeds a threshold t. More formally the average of:

$$\#Det_t(I_A, I_B) = |\{f_A^i \in F_A : IoU_{max}(f_A^i, F_B) > t\}| \quad (3)$$

To evaluate the repeatability of a feature detector, we compute a curve showing the average number detections $\#Det_t$ for a set of thresholds t.

Since we do not apply the normalization of [12] on the feature size, there is a bias towards large regions in the *IoU*. On the one hand, this effect is intended to obtain features whose appearance is less affected by the environmental changes, on the other hand we need to prevent that feature detectors gain significant performance just due to their region size. Therefore, we evaluate the amount of overlap of features between *corresponding* scenes relative to the average overlap when comparing *random pairs* of scenes. The resulting curves in Fig. 3 show the average rate of detections per corresponding image that exceeds the expected rate of "detections" in random images. Thus, artificial enhancement of the pure overlap performance that is not related to real image features also enhances the overlap on random scene pairs and does not affect the measure. And in contrast, if larger regions really increase the rate on common world points compared to random image pairs, than this results in an improved performance measure.

B. The Set of Compared Local Region Detectors

There are different algorithmic approaches to provide the spatial image support for place recognition landmarks. Keypoint detectors are a common choice for local region detectors. However, our prior experiments in [2] have shown that e.g. the keypoint based FAB-MAP [7] system has severe problems with seasonal changes. We include keypoints from Scale Invariant Feature Transform[2] (SIFT) [6] and Maximally Stable Extremal Regions[2] (MSER) [13] in this comparison. MSER finds regions that remain stable over a certain number of thresholds. These regions are invariant under affine image transformations and monotonic transformation of image intensities.

SIFT features are the scale space maxima and minima obtained by a difference of Gaussians (DoG) function on smoothed and resampled images. We want to clearly point out that we use the SIFT *detector* (the DoG scale space maxima detector) but not the SIFT *descriptor*. No single SIFT *descriptor* has been computed during this work. The local region around the SIFT scale space maxima is computed based on the scale at which the maximum is detected and a magnification factor M. To evaluate the influence of this magnification factor we evaluate the set of $M \in \{1, 6, 10, 15, 20\}$ yielding SIFT1, SIFT6 and so on. The receptive field of a SIFT *descriptor* is typically at magnification 6 [6].

A main task of the feature detector is to provide spatial image support for the subsequent description step. Similar to the region growing step in MSER keypoints, superpixel segmentation algorithms provide a broad spectrum of methods to provide such spatial image support. Superpixel segmentations are an oversegmentation of an image or - seen the other way around, a perceptual grouping of pixels. Oversegmention algorithms were not specifically designed to select salient features. However, the detected regions are somewhere located between small keypoints and the spatial image support for holistic image descriptors. Hence we include them in the set of promising "feature detectors". In prior work [14] we evaluated the stability of such algorithms across image changes. We found Felzenszwalb-Huttenlocher segmentation (FH)[3] [15], Quickshift Segmentation[2] (QS) [16] and Edge Augmented Mean Shift[4] (EAMS) [17] to be sufficiently stable to be promising approaches for providing spatial image support for across seasonal landmarks. FH is a graph based segmentation algorithm that incorporates the evidence of a boundary between two segments in a greedy algorithm that also grants global properties. EAMS combines a mean shift segmentation and a confidence based edge detector. QS is a graph based variant of medoid shift.

During superpixel segmentation, there is no criteria involved to foster the creation of *descriptive* segments that are repeatedly detected across severe appearance changes. Therefore, we use superpixels in three different ways:

1) Superpixel segmentations directly
2) A segment soup combining multiple segmentations of the same image
3) Randomized Prim's (RP)[5] [18], an object proposal algorithm that uses additional criteria to combine small scale superpixels to object candidates

Object proposal algorithms were designed as preprocessing steps for object detection to reduce the number of sliding windows that are classified to a much smaller set of promising candidates. RP uses a greedy algorithm for computing sets of FH superpixels that are likely to occur together. We further include Objectness[6] [19] as a second object proposal algorithm: It is one of the earliest approaches and samples and ranks a large number of windows per image according to their likelihood of containing an object. This likelihood is based on multiple cues derived from saliency, edges, superpixels, colour and location.

C. Experimental Results on Detector Repeatability

We use a 100 images subset of the training dataset from [2] as basis for the landmark repeatability evaluation. The original dataset consists of image pairs showing the same scene, from the same viewpoint at different seasons. To prevent in particular superpixel segmentation algorithms to exploit the image alignment, we crop the first image on the left side and the second image on the right side. This is equivalent to a shift of about 12.5 % of the image width. We configured the detectors to create about 50 features per image. Parameters that were changed from the defaults are given in Table I.

Fig. 3 shows the rate of features for which there is a matching with an overlap larger or equal the threshold t. As described above, these curves are normalized by the average rate of detections for random image pairs. I.e. a detector that just randomly samples regions would appear as a horizontal line at the abscissa. Please keep in mind that these detections are between images from different seasons including severe appearance changes. As a reading example: The upper most

[2]We used the implementation by http://www.vlfeat.org/

[3]http://www.cs.brown.edu/~pff/segment/

[4]http://www.wisdom.weizmann.ac.il/~bagon/matlab.html

[5]http://www.vision.ee.ethz.ch/~smanenfr/rp/index.html

[6]http://groups.inf.ed.ac.uk/calvin/objectness/

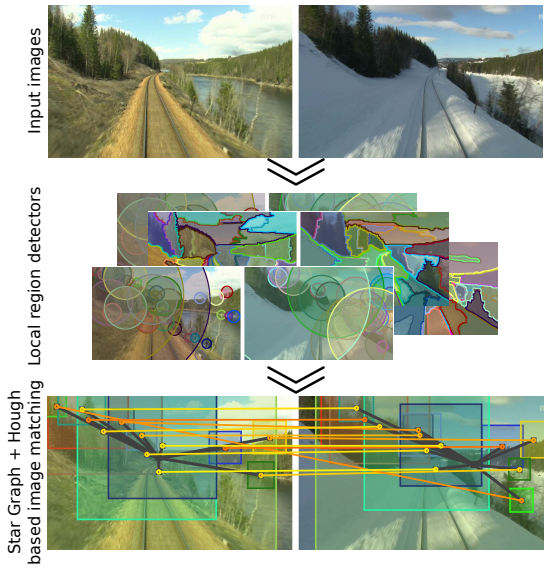

Fig. 2. Input to the matching procedure are images showing severe appearance changes and/or different viewpoints. The matching procedure can be used with a variety of possible local regions detectors as is illustrated in the middle row. Each landmark is described by a CNN descriptor that is used to find one-to-one matchings between two images as shown by the connections between the images in the bottom row. To incorporate the global landmark configuration in the image, the landmarks are arranged in a star graph model shown in dark grey. To cope with false matchings (e.g. the very diagonal matching), the final image similarities are computed by voting in a Hough scheme for the centre of the star-graph model.

SIFT	octaves=5, firstOctave=1, peakThresh=6.0
MSER	minDiversity=0.8, maxVariation=0.25, delta=5, minArea=200
FH	k=1, minSize=1500
EAMS	hs=8, hr=4, minSize=2000
QS	Ratio=0.5, kernelsize=2, maxDist=40
FH soup	3 FH segmentations: k=1, minSize $\in\{2500, 5000, 10000\}$
Objectness	n=50
RP	q=200

TABLE I. PARAMETERS THAT ARE DIFFERENT FROM THE DEFAULTS.

point of the dark blue curve indicates that FH creates for corresponding images on average 15 % detections with overlap ≥ 0.3 more than for random image pairs. For application as spatial image support for feature descriptors, high numbers of detections at high overlap are requested.

As expected, using directly the scale space maxima in SIFT1 provides high rates at small overlap. The best performance is obtained from SIFT with larger magnification factors. The typical choice for the SIFT *descriptor*, $M = 6$ shows the highest number of detection in the range of overlaps between 0.15 to 0.4. Increasing the magnification results in a higher number of detections for higher overlaps, at the cost of considerably less detections at mid-low overlaps. Please keep in mind, that this is already normalized by comparison to random image comparisons, thus we can hope to gain a real benefit for larger overlap values. Such dependencies on the region size vanish in the metric of Mikolajczyk et al. [12].

The MSER features vary more strongly in the presence of severe appearance changes as they occur in this spring-winter dataset, thus they perform inferior to SIFT. The basic oversegmentation features (FH,EAMS,QS) show peaks in the midrange of the overlap spectrum with FH providing the best results except for low overlaps were EAMS provides more features. These segmentation algorithms do not allow overlapping features, they are disjunct and distinct image

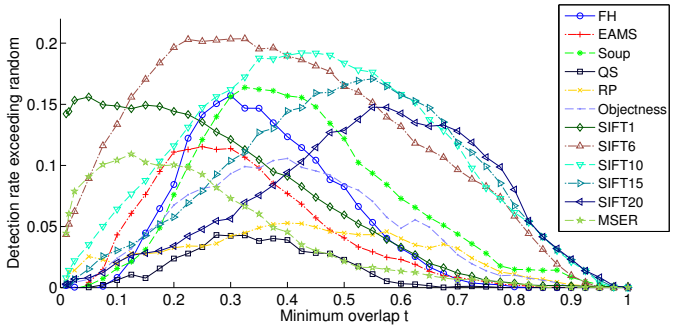

Fig. 3. Results on the overlap experiment using 50 landmarks per image.

partitions. In contrast, the overlapping segments in the FH soup (created from three FH segmentations, in total with 50 features) may overlap. This provides a variety on possible spatial image supports around a certain image point and seems to compensate some appearance changes. It can be seen, that FH soup clearly provides the largest number of features with considerable overlap from the set of segmentation based algorithms The object proposal methods Objectness and RP perform inferior. They have been designed to provide large number of features (e.g. 1,000). For Objectness we use the 50 features with the highest objectness-score. For RP we force the features to be as different as possible to reduce their number. This may prevent them from providing better repeatable features.

From this evaluation, the SIFT *detector* with a magnification factor 6 (as it is also typically used for the SIFT *descriptor*) can be considered to perform best. SIFT with higher magnification factors of 10 or 15 also seem to provide promising trade-offs between increased number of high overlap features without to much performance loss in the mid range. However, increasing the size of a local region detector is expected to increase the sensitivity against viewpoint changes and partial occlusions. We therefore present a novel approach for combination of the inflated SIFT detector with spatial image support from superpixels after the introduction of the overall matching methodology in the following section.

IV. THE PROPOSED LOCAL REGION DETECTOR + CNN LANDMARK BASED PLACE RECOGNITION APPROACH

In this paper we embrace the idea of CNN features as descriptors for place recognition in changing environments. In [3] Sünderhauf et al. propose to use the stacked output of a single CNN layer as descriptor for the whole image. They investigated several layers and found the lower convolutional layers to be robust against seasonal changes but sensitive to shifts, e.g. induced by viewpoint changes or rotations. To address the problems of the CNN based holistic image descriptor in the presence of viewpoint changes, we propose to compute a set of local CNN descriptors, one for each spatial image support provided by local region detector such as those compared in the previous section. To get the set of landmarks for an image, we process the following steps (cf. Fig. 2):

1) Run the local region detector of your choice on the image. For our experiments, again, we use 50 landmarks per image.
2) Compute the CNN descriptor for the image region inside the bounding box around each landmark. The CNN descriptor is the vectorized output of the third convolutional (*conv3*) layer of the VGG-M network[2].

This is the layer that showed best performance in across season matching and that is also used for the holistic image descriptor.

3) We arrange all local landmarks in a simple but flexible *star graph model* similar to the Implicit Shape Models of [5] to incorporate the landmarks' arrangement. Therefore, we compute the relative location of each landmark to the centre of the landmark set.

For comparing two images based on their landmarks, we compute the pairwise similarities from cosine distance between the landmarks' CNN feature vectors. These similarities are used as weights in a Hough scheme [20] to vote for the horizontal and vertical shift between the images' star graph centres. The similarity of the images is the maximum value in the resulting 2d Hough space. This way, we can incorporate the landmark arrangement in a scheme that can cope with small numbers of features including outliers and that is robust to variations in the exact landmark locations (e.g. compared to Fundamental matrix estimation). To reduce the number of outlier matchings between individual landmarks, we also incorporate a left-right check to select features that are included in the Hough voting. An implementation of the proposed matching procedure is provided on our website.[1]

As stated above, increasing the size of landmarks is expected to increase their sensitivity to viewpoint changes and partial occlusions. We propose an optional additional step: a combination of the inflated scale space extrema from the SIFT detector with spatial support from superpixel segmentations. Superpixel segmentations are designed to preserve object boundaries. Navarro et al. [21] propose to split SIFT regions into foreground and background based on a superpixel segmentation and describe only the foreground part. We extend this approach by computing a weight matrix from a superpixel segment soup to weight the importance of each dimension of the CNN descriptor. Again, we compute the segment soup using FH and vary the following parameters to create the soup: $k \in \{1, 10, 100, 1000\}$ and $minSize \in \{2500, 5000, 10000\}$. For lower layers of CNN features (as is the used conv3 layer), each dimension has a limited receptive field in the described image region [22]. We create a weight matrix of the spatial resolution of the used CNN layer (e.g. 13×13 for the used conv3 layer). Since the descriptors are compared using cosine distance, the normalized weight matrix values can be directly multiplied with the corresponding dimensions. The weight values are computed as follows: For each segmentation mask from the segment soup, we compute the IoU with the landmark and increase the weight of all pixels covered by this mask by the IoU value. This way, pixels that are contained in segments that are well aligned with the landmark will gather higher values than pixel that are located in segments that overflood the landmark and are thus likely to belong to the background.

V. Place Recognition Experiments

We use the challenging Nordland test dataset from [2] for our experiments. This dataset provides images of a train ride in Norway from all four seasons. We use each tenth frame of the full 700+ km ride. Example images can be seen in Fig. 1 and 2. Results of other approaches on this dataset can be found in [2], [3] and [4]. E.g. the SIFT (detector and descriptor) based FAB-MAP fails on this dataset. The best performing existing method is the holistic CNN feature based matching from [3]. The original Nordland images are almost

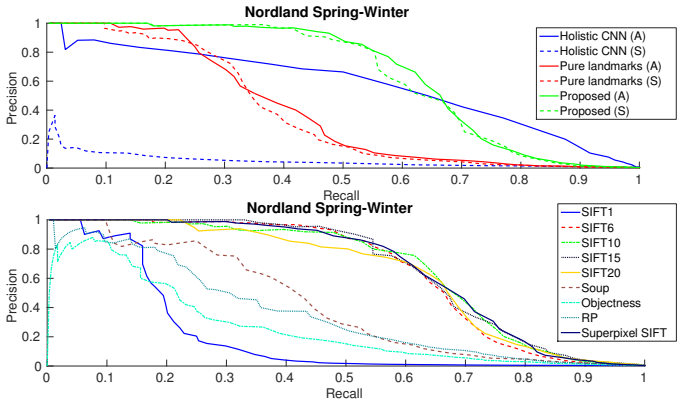

Fig. 4. Place recognition results. Top: he proposed landmarks (SIFT6 + CNN) stand alone (*Pure landmarks*) and in combination with the *Proposed* star graph and Hough based image matching approach in comparison with the holistic descriptor on aligned images (A) and shifted image pairs (S). Bottom: Results of various local region detectors in combination with the proposed image matching approach on the aligned Nordland images.

perfectly pixel aligned. While these are perfect conditions for the holistic CNN image descriptor, they are not realistic in practical scenarios. Following [3], we additionally shift (and subsequently crop) one image set to the left (spring) and the other image set to the right (winter) by 10 % of the image width. The effect of the 10% shift can roughly be compared to the rotation of the used camera of about 5 degrees (or a slightly different pose). We create precision-recall curves by varying a threshold on the image distance and classifying each image comparison as positive or negative matching.

Fig. 4 shows the benefit of the presented approach using SIFT6 compared to the best existing holistic matching approach. The blue curves show matching based on the holistic CNN image descriptor as described in [3]. The holistic approach fails in the presence of the artificial image shift (S). The red curves show the performance when using the proposed landmarks directly without the star graph model and Hough based approach. Therefore, the similarity between two images with landmark sets A and B is directly computed from the distance of all left-right matchings and the total number of landmarks:

$$sim(A, B) = \frac{1}{\sqrt{|A| \cdot |B|}} \sum_{\{a,b\} \in matches(A,B)} 1 - dist(a, b) \quad (4)$$

The green curves show that additional incorporation of the proposed star graph model and Hough based image matching approach further increases the place recognition performance significantly. The bottom part of Fig. 4 shows results for a subset of the previously evaluated detectors. It can be seen that the general performance on the proposed measure of Sec. III-C is resembled: As expected, SIFT1 without rescaling performs inferior and SIFT 6 to 15 show the best performance. The difference in the relation of the two object proposal algorithms might be due to the missing shift that was induced in the overlap evaluation dataset but is missing in the aligned images setup. Superpixel SIFT is the proposed combination of a SIFT10 with a descriptor reweighting based on the IoU with a superpixel soup. For the presented experiment on the aligned Nordland images there is no significant influence of this additional effort visible.

To evaluate possible negative influence of the resizing of the SIFT regions, we run an additional experiment on the

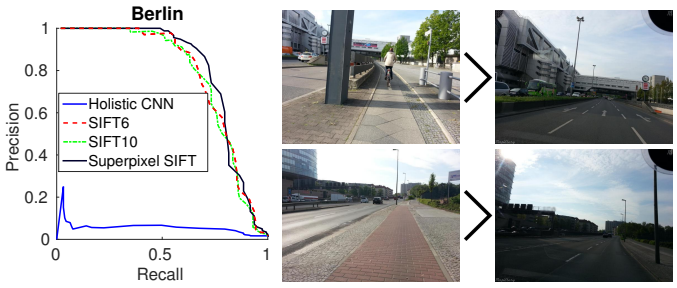

Fig. 5. Example images and place recognition results on the Berlin dataset.

Berlin Halenseestrae Mapillary[7] dataset. It provides two short image sequences (157 and 67 frames, 3 km), one from a bicyclist on a bike lane and the second from a car. Example images showing the severe viewpoint changes can be seen in Fig. 5 together with place recognition results. Due to the similarity of consecutive frames we allow matchings to the one provided ground-truth image and the two neighboured frames. As expected the holistic descriptor fails, while the proposed approach provides reasonable performance. Further, there is a small benefit from using the additional superpixel based reweighting for this setup.

VI. Conclusions

I this paper we proposed and evaluated the combination of local region detectors and CNN descriptors as landmarks for place recognition in changing environments. The contributions of this paper are twofold: first we compared several region detectors based on their ability to provide image regions that show large overlap across severe image changes as they appear between different seasons. We want to point out, that in terms of measuring the repeatability of feature detectors, we do not want to replace the methodology presented in [12], but we want to extend the insights on the importance of feature sizes in particular in the presence of changing environments.

Second, we proposed an image matching approach that can be combined with a broad range of local region detectors including those without stable midpoints like segmentation based approaches or object proposals. The presented preliminary experiments showed the benefit of the proposed landmarks (local regions detectors and CNN descriptors) and the additional benefit from the proposed image matching approach based on star graph models and Hough voting. We discussed the influence of the size of the used local regions and found indications that the scale space extrema detector used in the SIFT *detection* step can also be used with larger spatial image support than for the typical SIFT *descriptor*. We also demonstrated a potential benefit of the combination of such an increased region with the spatial image support of a superpixel soup.

This work presented several aspects that can be included in the design of a place recognition system ready for practical applications. In particular, a matching framework that can be combined with a broad range of existing and new region detectors, e.g. novel object proposal algorithms. Two main directions for future work will be the experimental validation beyond the preliminary results and work on practical aspects like runtime and confidence of image comparisons.

For the proposed evaluation criterion for local regions detectors it would interesting to see whether the normalization with random image pairs can be enhanced or replaced with a more analytical methodology that takes the feature size and distribution in the image into account.

[7] http://www.mapillary.com. We want to thank Niko Sünderhauf for this dataset.

References

[1] M. Milford and G. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights." in *Int. Conf. on Robotics and Automation (ICRA)*, 2012.

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

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

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

[5] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Workshop on Statistical Learning in Computer Vision (ECCV workshop)*, 2004.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, 2004.

[7] M. Cummins and P. Newman, "FAB-MAP:probabilistic localization and mapping in the space of appearance," *Int. J. Rob. Res.*, jun 2008.

[8] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *Int. Conf. on Comp. Vision (ICCV)*, 2003.

[9] W. Churchill and P. Newman, " Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation." in *Int. Conf. on Robotics and Automation (ICRA)*, 2012.

[10] C. Valgren and A. Lilienthal, "SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments," *Robot. Auton. Syst.*, no. 2, 2010.

[11] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," *CoRR*, vol. abs/1403.6382, 2014. [Online]. Available: http://arxiv.org/abs/1403.6382

[12] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *Int. J. Comput. Vision*, vol. 65, pp. 43–72, 2005.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. of British Machine Vision Conference (BMVC)*, 2002.

[14] P. Neubert and P. Protzel, "Evaluating Superpixels in Video: Metrics Beyond Figure-Ground Segmentation." in *Proc. of British Machine Vision Conference (BMVC)*, 2013.

[15] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, 2004.

[16] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *European Conf. on Computer Vision (ECCV)*, 2008.

[17] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. on Pat. Anal. and Mach. Intel.*, vol. 24, 2002.

[18] S. Manén, M. Guillaumin, and L. Van Gool, "Prime Object Proposals with Randomized Prim's Algorithm," in *Proc of. Interna. Conf. on Computer Vision (ICCV)*, 2013.

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

[20] P. Hough, "Method and Means for Recognizing Complex Patterns," U.S. Patent 3.069.654, Dec. 1962.

[21] F. Navarro, M. Escudero-Violo, and J. Bescos, "Sp-sift: enhancing sift discrimination via super-pixel-based foreground-background segregation," *Electronics Letters*, vol. 50, no. 4, pp. 272–274, February 2014.

[22] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *CoRR*, vol. abs/1412.6856, 2014. [Online]. Available: http://arxiv.org/abs/1412.6856