Journal of Robotics and Autonomous Systems (RAS), Volume 69, July 2015, Pages 15-27
DOI: http://dx.doi.org/10.1016/j.robot.2014.08.005

Superpixel-based Appearance Change Prediction for Long-Term Navigation Across Seasons

Peer Neubert, Niko Sünderhauf, Peter Protzel

***Abstract*— Changing environments pose a serious problem to current robotic systems aiming at long term operation under varying seasons or local weather conditions. This paper build on our previous work where we propose to learn to *predict* the changes in an environment. Our key insight is that the occurring scene changes are in part systematic, repeatable and therefore predictable. The goal of our work is to support existing approaches to place recognition by learning how the visual appearance of an environment changes over time and by using this learned knowledge to predict its appearance under different environmental conditions. We describe the general idea of appearance change prediction (ACP) and investigate properties of our novel implementation based on vocabularies of superpixels (SP-ACP). Our previous work showed that the proposed approach significantly improves the performance of SeqSLAM and BRIEF-Gist for place recognition on a subset of the Nordland dataset under extremely different environmental conditions in summer and winter. This paper deepens the understanding of the proposed SP-ACP system and evaluates the influence of its parameters. We present the results of a large-scale experiment on the complete 10 hour Nordland dataset and appearance change predictions between different combinations of seasons.**

I. Introduction

Long term navigation in changing environments is one of the major challenges in robotics today. 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 depending on the current season, weather conditions or the 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.

The problem has recently been addressed by few authors, but so far no congruent solution has been proposed. Milford and Wyeth [17] proposed to increase the place recognition robustness by matching *sequences* of images instead of single images and achieved impressive results on two across-seasons datasets. Exploring into a different direction, Churchill and Newman [5] proposed to accept that a single place can have a variety of appearances. Their conclusion was that instead of attempting to match different appearances across seasons or severe weather changes, different *experiences* should be remembered for each place, where each experience covers exactly one appearance. Both suggested approaches can be understood as the extreme ends of a spectrum of approaches that spans between interpreting changes as individual experiences of a single place on one hand and increasing the robustness of the matching against appearance changes on the other hand. Our work presented in the following is orthogonal to this spectrum.

The authors are with the Department of Electrical Engineering and Information Technology, Chemnitz University of Technology, 09111 Chemnitz, Germany. Contact: peer.neubert@etit.tu-chemnitz.de
Website: http://www.tu-chemnitz.de/etit/proaut

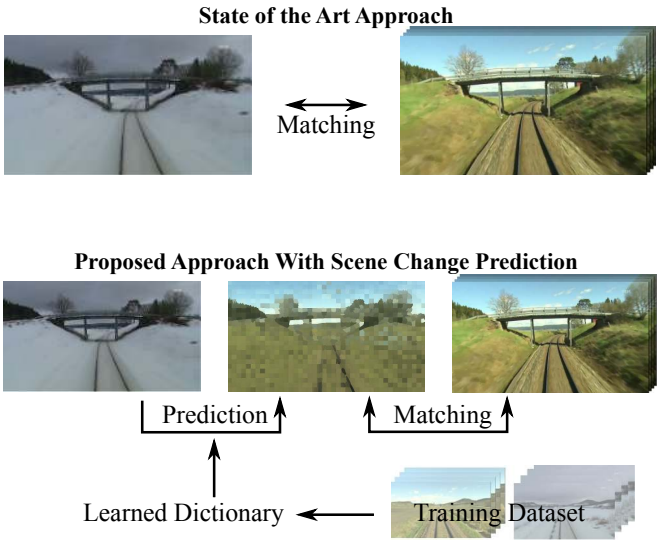

Fig. 1. State of the art approaches to place recognition will attempt to directly match two scenes even if they have been observed under extremely different environmental conditions. This is prone to error and leads to bad recognition results. Instead, we propose to *predict* how the query scene (the winter image) would appear under the same environmental conditions as the database images (summer). This prediction process uses a dictionary that exploits the systematic nature of the seasonal changes and is learned from training data.

What current approaches to place recognition (and environmental perception in general) lack, is the ability to *reason* about the occurring changes in the environment. Most approaches try to merely *cope* with them by developing change-invariant descriptors or matching methods. Potentially more promising is to develop a system that can *learn* to *predict* certain systematic changes (e.g. day-night cycles, weather and seasonal effects, re-occurring patterns in environments where robots interact with humans) and to infer further information from these changes. Doing so without being forced to explicitly know about the *semantics* of objects in the environment is in the focus of our research and the topic of this paper.

Fig. 1 illustrates the core idea of our work and how it compares to the current state of the art place recognition algorithms. Suppose a robot re-visits a place under extremely

different environmental conditions. For example, an environment was first experienced in summer and is later re-visited in winter time. Most certainly, the visual appearance has undergone extreme changes. Despite that, state of the art approaches would attempt to match the currently seen winter image against the stored summer images.

Instead, we propose to *predict* how the current scene would appear under the same environmental conditions as the stored past representations, before attempting to match against the database. That is, when we attempt to match against a database of summer images but are in winter time now, we predict how the currently observed winter scene would appear in summer time or vice versa.

The result of this prediction process is a synthesized summer image that preserves the structure of the original scene and is close in appearance to the corresponding original summer scene. This prediction can be understood as *translating* the image from a winter vocabulary into a summer vocabulary or from winter language into summer language. As is the case with translations of speech or written text, some details will be lost in the process, but the overall *idea*, i.e. the gist of the scene will be preserved. Sticking to the analogy, the error rate of a translator will drop with experience. The same can be expected of our proposed system: It is dependent on training data, and the more and the better training data is gets, the better can it learn to predict how a scene changes over time or even across seasons.

This paper build upon our previous work [18] where we introduced the novel idea of predicting extreme scene changes across seasons to aid place recognition for the first time. We prove the feasibility of our idea and describe an implementation based on superpixel vocabularies. We demonstrate how we can predict the appearance of natural scenes across winter and summer time, as illustrated in Fig. 1. By applying this approach, we are able to significantly improve the place recognition performance of SeqSLAM [17] and BRIEF-Gist [21] on the new, publicly available large-scale Nordland dataset [20] that traverses an environment in winter and summer under extremely different environmental conditions. While the first results we reported in [18] were based on a small subset of the Nordland dataset, this paper presents new results on the *complete* Nordland track. We furthermore evaluate predictions between different combinations of seasons. An extensive evaluation of important parameters deepens our understanding of the proposed prediction system and its parameters.

In the following section, we put the proposed prediction system in the context of related work, before we describe its algorithmic steps in section III. Section IV introduces the Nordland dataset we used for training, validation and testing. The results section V presents comprehensive place recognition experiments on this dataset using FAB-MAP, BRIEF-Gist and SeqSLAM in combination with the proposed SP-ACP system. The paper is concluded by a discussion of limitations of the current system and directions for future work in section VI. Additional information and videos can be obtained from our project website[1].

II. Related Work

The related work is threefold. First we give a short review of the work on visual place recognition in changing environments, followed by methods on how to deal with changing environments on the mapping side, finally we present the relation of our approach to the texture transfer and image analogy ideas published in computer graphics.

A. Approaches for place recognition in changing environments

Traditionally, visual place recognition is either based on matching local features (like SIFT or SURF keypoints), bags of visual words (like FAB-MAP [6]), global image descriptors (like GIST [24]), or a combination. While there is a large body of research on visual place recognition in static scenes or scenes with few moving objects, only recently attempts were made to extend the recognition capabilities to changing environments, e.g. to achieve across-season matchings. So far four directions exist how such changing environments can be dealt with:

1) Using standard approaches and hope for the best
2) Increase robustness by matching image sequences
3) Switching to wavelengths other than visible light
4) Searching for seasonally invariant features

In the following we give a short overview of attempts in each direction.

1) Using standard approaches and hope for the best: Based on local keypoint features, Valgren and Lilienthal [25] show high recognition rates on single image matching of five places across seasons. Their approach uses U-SURF keypoints and descriptors on omnidirecional images. They conclude that high-resolution omnidirectional images and additional constraints on the matched keypoints (epipolar geometry and reciprocal matchings) are necessary. Unfortunately, it remains unclear what portion of matchings are on seasonally invariant objects (like building facades) and how this approach generalizes to larger datasets. Keypoint based approaches in changing environments have to rely on the detection of keypoints on objects that vary strongly in their appearance. To overcome this shortcoming, a fixed distribution of keypoints can be used (e.g. a keypoint grid). Sift Flow[14] computes a SIFT descriptor at each pixel and matching is based on local and global constraints. While they show impressive results on scene alignment under strongly varying conditions, this approach has not yet been used for across season place recognition. Glover et al. [8] present a combination of the advanced local feature recognition system FAB-MAP [6] and the biologically inspired SLAM approach RatSLAM [16] based on pose cell filtering and experience mapping. RatSLAM is robust to false-positive loop closures from the image processing front-end and integrates matching information over time. The hybrid FAB-MAP + RatSLAM

[1] http://www.tu-chemnitz.de/etit/proaut/forschung/acp.html

system has shown that mapping in challenging outdoor conditions with variances due to illumination and structure is possible. However, the authors conclude that the SURF features on which it is based, are too variable under those varying conditions to form a truly reusable map.

2) Increase robustness by matching image sequences: The pose cell filtering of RatSLAM is a step towards using sequences for matching. In their subsequent work, the RatSLAM authors presented SeqSLAM [17] that builds upon a lightweight 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. Although their system is limited to constant velocity motion, it represents the state of the art for matching under changing conditions. Badino et al. [2] 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.

3) Switching to wavelengths other than visible light: Maddern and Vidas [15] combine visible and long-wave infrared imaging for place recognition through a day-night-cycle. Their system is based on FAB-MAP and combines words of SURF features from the visible and infrared images (using two separate vocabularies). They find the combination of both modalities to give the best results: infrared is more robust to extreme changes while the visible modality provides better recall during day. They present preliminary results on data of a 1.5 km track traversed several times during a single day-night cycle.

4) Searching for seasonally invariant features: He et al. [10] learn an intermediate representation of images such that the distance of two images in this intermediate representation reflects the distance between the places in the world, where these images were taken. The intermediate representation is a vector of weighted SIFT feature prototypes. Since they train their system on summer and winter images, they search for a set of SIFT features that are suitable for place recognition under this seasonal change. Their approach still relies on the extraction of local keypoints on the same world object under the seasonal change. Zhang and Kosecka [26] focus on recognizing buildings in images. They use a hierarchical matching scheme based on localized color histograms and SIFT features to search for buildings in an image database. While they did not explicitly design their system for place recognition across seasons, their test data (ZuBud) covers different weather conditions and seasons, thus building facades could serve as seasonally invariant landmarks.

B. Organizing the Changes in a Map

Changing environments are challenging for visual place recognition systems. But they are also a challenge for the mapping side of the problem. Churchill and Newman [5] present a mapping system based on a plastic map, a composite representation of multiple experiences connected in a relative framework. Each experience handles a sequence of images, motion and 3D feature data. Multiple localizers match the current frame to stored experiences. Several experiences can be active at once, when they represent the same place. The complexity of the plastic map varies according to the amount of variation in the scene. They present results in changing lighting and weather conditions over a three month period. For pose graph SLAM, Biber and Duckett [4] showed that the map grows unbounded in time, even for small environments that are repeatedly traversed. Johannsson et al. [12] proposed the reduced pose graph that reuses already existing poses in previously mapped areas and incorporates new measurements as new constraints between existing poses. This can be used if the place recognition frontend can match the poses. Konolige and Bowman [13] present a mapping system based on a skeleton graph of keyframes from a visual odometry system. Views of keyframes are updated and deleted to preserve view diversity while limiting their number. They showed their system to handle changing light conditions in an office environment.

In summary, being able to associate places despite severe changes in their appearance is advantageous to the mapping process since the rate at which new experiences [5], poses [12], or views [13] have to be introduced to the map can be reduced.

C. Correspondence to the Texture Transfer and Image Analogy Problems

The idea to predict images from training examples has some relations to two other image processing tasks:

The texture transfer problem [7]: Given two images A_S, A_W and a correspondence map C that relates parts of A_S to parts of A_W, synthesize the first image with the texture of the second. C typically depends on image intensity, color, local image orientation or other derived quantities.

The image analogy problem [11]: Given an image pair (A_S, A_W) and a query image B_S, compute a new "analogous" image B_W that relates to B_S in the same way as A_W to A_S.

Speaking in the context of predicting image change across seasons: A_S, A_W are given summer (S) and winter (W) training images and we learn to synthesize a new winter image B_W given a new summer image B_S or vice versa. The approaches of Efros and Freeman [7] and Hertzmann et al. [11] create visually appealing results but have not yet been used in context of place recognition. They focus on using single image pairs instead of large collections of training data. Nevertheless, such approaches could be used to improve the visual coherence of the images predicted by the proposed prediction framework.

III. SP-ACP: Learning to Predict Scene Changes Across Seasons

In this section of our paper we explore how the changing appearance of a scene across different environmental conditions can be predicted. Throughout the remainder of this section these changing environmental conditions will be summer and winter. However, the concepts described in the

following can of course be applied to other sets of contrasting conditions such as day/night or weather conditions like sunny/rainy etc.

How can the severe changes in appearance a landscape undergoes between winter and summer be learned and predicted? The underlying idea of our approach is that the appearance change of the whole image is the result of the appearance change of its parts. If we had an idea of the behavior of each part, we could predict the whole image. Instead of trying to recover semantic information about the image parts and model their behavior explicitly, we make the assumption that similarly appearing parts change their appearance in a similar way. While this is for sure not always true, it seems to hold for many practical situations (e.g. changing color of the sky from sunny day light to dawn, appearance of a meadow in summer to its snow-covered winter counterpart). This idea can be extended to groups of parts, incorporating their mutual relationships.

We use superpixels (see section III-D.1 for details) as image parts and cluster them to vocabularies using a descriptor (III-D.2). To predict how the appearance of a scene changes between summer and winter, we first conduct a learning phase on training data (III-A) which comprises scenes observed under both summer and winter conditions. In the subsequent prediction phase (III-C), the appearance of a new image seen under one of the conditions is predicted as it would be observed under the other viewing condition.

A. Learning a Vocabulary for Summer and Winter

During the training phase we have to learn a vocabulary for each viewing condition and a dictionary to translate between them. In a scenario with two viewing conditions (e.g. summer and winter), the input to the training are images of the same scenes under both viewing conditions and known associations between pixels corresponding to the same world point. Obviously the best case would be perfectly aligned pairs of images. This requirement for almost pixel-accurately aligned training images is clearly the major limitation of our current system. The Nordland dataset discussed in Section IV fulfills these needs.

Fig. 2 illustrates the training phase. Each image is segmented into superpixels and a descriptor for each superpixel is computed. The set of descriptors for each viewing condition is clustered to a vocabulary using hierarchical k-means. Each cluster center becomes a word in this visual vocabulary. The descriptors and the average appearance of each word (the word patch) are stored for later synthesizing of new images. For our experiments, we learned 10,000 words for each vocabulary.

B. Learning a Dictionary to Translate between Vocabularies

The learned visual vocabularies for both summer and winter conditions are able to express a typical scene from their respective season. The next step is learning a dictionary that allows translating between both vocabularies. This is illustrated in the lower part of Fig. 2.

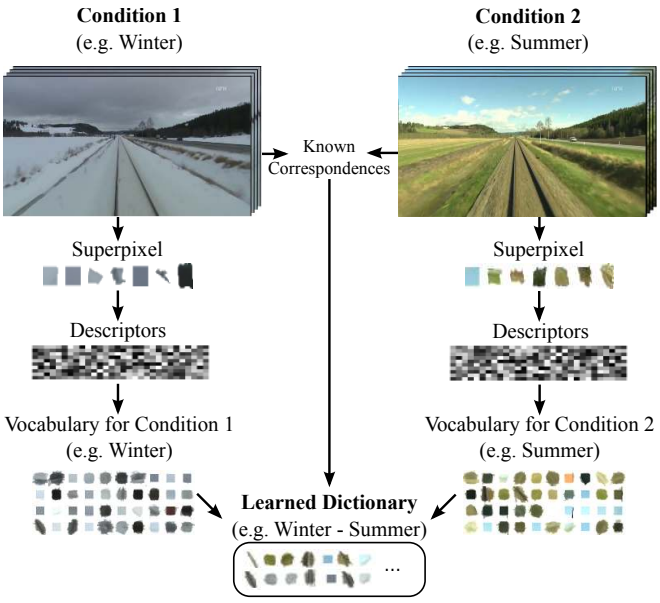

Fig. 2. SP-ACP learning a dictionary between images under different environmental conditions (e.g. winter and summer). The images are first segmented into superpixels and a descriptor is calculated for each superpixel. These descriptors are then clustered to obtain a vocabulary of visual words for each condition. In a final step, a dictionary that translates between both vocabularies is learned. This can be done due to the known pixel-accurate correspondences between the input images.

Since the images from the training dataset are aligned, we can determine how single words behave when the environmental conditions change. By overlaying the two aligned images from both summer and winter conditions, every pixel is associated with two words, one from the winter and another from the summer vocabulary. For each combination of words from the summer and winter vocabulary we can then count how often they have been associated to the same pixel coordinates.

This process is repeated for every pair of corresponding images in the training dataset, step-by-step building a distribution over the occurring translations between words from one vocabulary into the other. The final dictionary can be compiled by either storing the full distribution or ignoring it and using a winner-takes-all scheme that stores only the transition that occurs most often. The experimental results of section V will compare both approaches.

C. Predicting Image Appearances Across Seasons

Fig. 3 illustrates how we can use the learned vocabularies and the dictionary to predict the appearance of a query image across different environmental conditions.

The query image is segmented into superpixels and a descriptor for each superpixel is computed. Using this descriptor, a word from the vocabulary corresponding to the current environmental conditions (e.g. winter) is assigned to each superpixel. The learned dictionary between the query conditions and the target conditions (e.g. winter-summer) is used to translate these words into words of the target vocabulary.

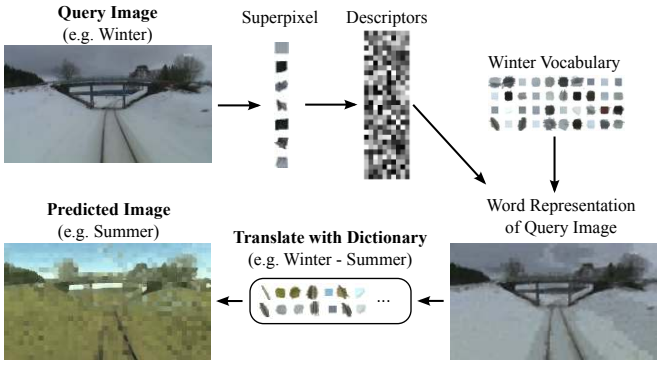

Fig. 3. SP-ACP predicting the appearance of a query image under different environmental conditions: How would the current winter scene appear in summer? The query image is first segmented into superpixels and a descriptor is calculated for each of these segments. With this descriptor each superpixel can be classified as one of the visual words from the vocabulary. This word image representation can then be translated into the vocabulary of the target scene (e.g. summer) through the dictionary learned during the training phase (see Fig. 9). The result of the process is a synthesized image that predicts the appearance of the winter query image in summer time.

Since the vocabularies also contain *word patches*, i.e. an expected appearance of each word, we can synthesize the predicted image based on the word associations from the dictionary and the spatial support given by the superpixel segmentation. Notice that when the dictionary provides the full distribution over possible translations for a word (as opposed to the winner-takes-all scheme), the resulting synthesized image patches are built by the weighted mean over all patches from the target words in the distribution. No further processing (e.g. as proposed by [7], [11]) is done to improve the appearance or smoothness of the resulting word images. Example word images and predictions are shown in Fig. 9.

D. Superpixel Segments and their Descriptors

As we have seen, superpixel segments play an important role in our proposed approach. They constitute the parts of the images and carry the information that is exploited for learning and predicting. We therefore want to briefly provide information on the used segmentation algorithm and descriptors.

1) Superpixels: Superpixels are the result of perceptual grouping of pixels or seen the other way around, the result of an image oversegmentation. In contrast to an object-ground segmentation, typically a superpixel segmentation divides the image into 25 - 2500 segments. Superpixels carry more information than pixels and align better with object edges than rectangular image patches. The term was coined in [19] and various algorithms for computation of superpixels exist.

In this work, we use a version of SLIC [1] to segment the image in 1,000 superpixels. SLIC performs a localized k-means to cluster pixels based on Lab-color values and pixel position in the image. The results are compact and regularly distributed superpixels following object boundaries. Example segmentations are shown in Fig. 4.

2) Superpixel Descriptor: Various descriptors for superpixels exist in the literature. Typically the descriptor includes various types of features combined with dimensionality reduction techniques. E.g. Tighe et al. [23] combine shape, location, texture (using SIFT) and color features. Barnard et al. [3] use 40 features, the descriptor of Gould et al. [9] even includes multiple color descriptors.

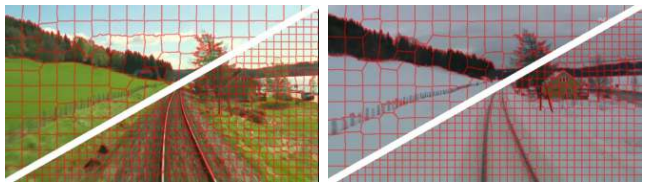

Fig. 4. Example superpixel segmentations. Two input images are segmented into 200 (top left triangles) and 1,000 (bottom right triangles) superpixels. Superpixel borders are shown red.

In the presented work we combine a color histogram in Lab color space (each channel with 10 bins) with an upright SURF descriptor (128 Byte) to capture texture. The SURF descriptor is computed over the entire superpixel, using the superpixel midpoint as keypoint. We additionally include the y-coordinate of the superpixel center. The influence of this additional information is evaluated in Fig. 10. We do not apply further dimensionality reduction.

IV. The Nordland Dataset

To test our proposed approach of appearance change prediction, we required a dataset where a camera traverses the same places under very different environmental conditions but under a similar viewing perspective. Ideally, the dataset should contain ground truth information, e.g. the corresponding scenes should be known.

The TV documentary "Nordlandsbanen – Minutt for Minutt" by the Norwegian Broadcasting Corporation NRK provides video footage of the 728 km long train ride between the cities of Trondheim and Bodø in north Norway. The complete 10 hour journey has been recorded from the perspective of the train driver four times, once in every season. Thus the dataset can be considered comprising a single 728 km long loop that is traversed four times. As illustrated in Fig. 5, there is an immense variation in the appearance of the landscape, featuring a complete snow cover in winter, fresh and green vegetation in spring and summer, as well as colored foliage in autumn.

In addition to the seasonal changes, different local weather conditions like sunshine, overcast skies, rain and snowfall are experienced on the long trip. Fig. 6 shows the altitude profile of the complete track and illustrates the high variance in appearance in a single season due to the different vegetation zones the train passes. Most of the journey leads through such natural scenery, but the train also passes through urban areas along the way and occasionally stops at train stations or signals.

The videos of the journey have been recorded at 25 fps with a resolution of 1920×1080 using a SonyXDcam with a Canon image stabilizing lens of type HJ15ex8.5B KRSE-V. GPS readings were recorded in conjunction with the video at 1 Hz. Both the videos and the GPS track are publicly

Fig. 5. The Nordland dataset consists of the video footage recorded on a 728 km long train ride in northern Norway. The journey was recorded four times, once in every season. Frame-accurate ground truth information makes this a perfect dataset to test place recognition algorithms under severe environmental changes. The four images above show the same place in winter, spring, summer and fall. Images licensed under Creative Commons (CC BY), Source: NRKbeta.no http://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/

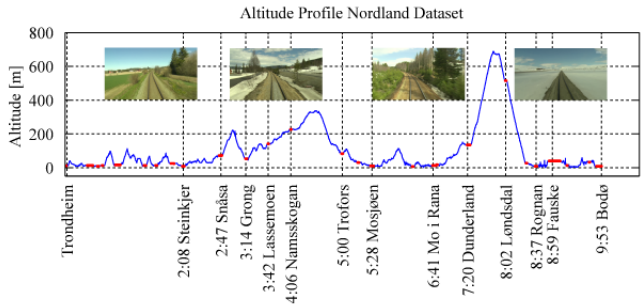

Fig. 6. On its 728 km long way from Trondheim to Bodø, the train moves through different vegetation zones and altitude levels. The plot shows the altitude profile along with typical images from the corresponding areas of the spring journey. Some of the stations and their arrival time in the videos are marked as well. The red dots indicate that the train stopped at stations or signals. The plot has been generated using the GPS data that is published along with the videos.

available online[2] under a Creative Commons licence (CC BY). The full-HD recordings have been time-synchronized by the TV company NRK such that the position of the train in an arbitrary frame from one video corresponds to the same frame in any of the other three videos. This was achieved by using the recorded GPS positions and interpolating the GPS measurements to 25 Hz to match the video frame rate. The dataset therefore meets the requirements for almost pixel-accurately aligned images during the training phase. Notice that the image alignment is not absolutely perfect due to GPS interpolation errors and the 25 Hz sampling speed while moving at high velocities. The misalignment between two corresponding images in the dataset is however rather minor and in most cases only visible in objects close to the train tracks, such as trees or poles.

[2] http://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/

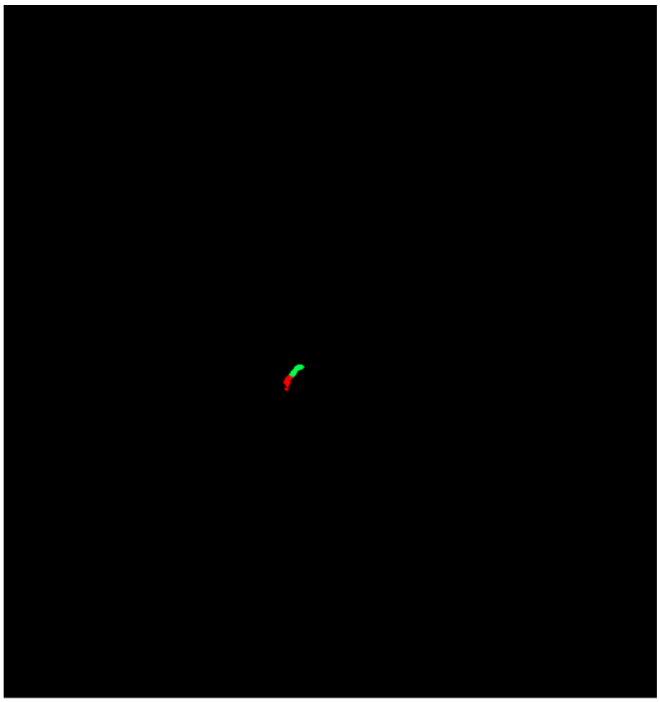

Fig. 7. GPS track of the complete Nordland dataset. The position of the training and validation data sets have been marked in red and green respectively. Map data ©2014 Google.

A. Training and Validation Datasets

For the training and evaluation experiments described in the following, we extracted 30 minutes from the spring and the winter videos at a frame rate of 2 Hz, starting approximately at 2 hours into the drive. From the four available videos, the spring video best resembled typical summer weather conditions at the chosen part of the track. Thus, most experiments are based on the spring and winter videos (otherwise it is explicitly mentioned). All images were resized to 854×480 pixels.

To extract this training and validation subset, we used the following *avconv* command on the Linux command line:

```
avconv -vsync vfr -r 2 -t 1800 -ss 02:02:00 -s 854x480
-i nordlandsbanen.winter.sync.1920x1080.h264.nrk.mp4
image-%05d.png
```

The first 900 frames (about 8 minutes) from this 30 minutes subset form the *training* dataset. This training dataset was used to learn the visual vocabulary for summer and winter and the dictionary to translate between both seasons. The last 2700 frames of the remaining 22 minutes of this video subset served as *validation* dataset to evaluate the influence of parameter settings of the proposed approach to scene change prediction. There are 200 unused frames between the training and validation dataset since the train stopped at a station during this time.

B. Test dataset

Besides the training and validation datasets, we also extracted a much larger *test* dataset that covers the complete spring and winter journey. For testing we subsampled the 10

TABLE I

OVERVIEW OF CONDUCTED EXPERIMENTS

Place Recog. Algorithm	Validation Dataset	Test Dataset
FAB-MAP	Section V-B	-
BRIEF-Gist	Section V-C	-
SeqSLAM	Sections V-D.2 and V-D.3	Section V-D.4

h videos at 0.1 frames per second using the following *avconv* command:

```
avconv -r 0.1 -vsync vfr -s 854x480
-i nordlandsbanen.winter.sync.1920x1080.h264.nrk.mp4
image-%05d.png
```

Fig. 7 illustrates the complete journey from Trondheim to Bodø and the position of the training and validation datasets.

V. EXPERIMENTS AND RESULTS

After the previous sections explained our proposed SP-ACP system and introduced the Nordland dataset, we can now describe the conducted experiments and their results.

We evaluate the proposed SP-ACP prediction system by using it as a preprocessing step to the existing place recognition algorithms FAB-MAP [6], BRIEF-GIST [21], and SeqSLAM [17]. For each of these three established approaches, we will compare the respective performance of

1) directly matching between images of different seasons, e.g. winter vs. spring
2) using the proposed SP-ACP to predict the changed appearance of one of the seasons and e.g. match a predicted winter image against the real winter images

We will calculate precision and recall and use apply the resulting F-score as the primary performance measure. Since modern SLAM systems do not have to rely on their place recognition front ends to operate at 100% precision anymore [22], the recall at 100% precision is used as a secondary performance indicator only. Experiments with all three mentioned algorithms will first be conducted on the *validation* dataset. We will use SeqSLAM to perform an analysis of the parameters of SP-ACP and determine the optimal values for these parameters. Finally, SP-ACP using these optimal settings is applied to the *test* dataset that covers the complete Nordland journey to demonstrate how place recognition in changing environments can benefit from our proposed appearance change prediction. Table I gives an overview of the experiments. Table II shows the default parameters. Deviations from these setting are indicated for each experiment.

A. Applying SP-ACP: Predicting Images of the Nordland Dataset

Using the training dataset (section IV-A) and the *training* procedure introduced in section III-A, we can learn vocabularies for spring and winter conditions and a translation (dictionary) between them. Fig. 8 shows example pairs of words from both vocabularies. For each spring word, we show the winter word with the highest impact on the translation (which is the one that would be applied in a WTA setup). Following the proposed *prediction* procedure of section III-C, we can use the learned vocabularies and dictionary to create a predicted winter image for a given spring image or vice versa. Example predictions are visualized in Fig. 9 and Fig. 14. To evaluate the benefit of the proposed prediction step for place recognition, we can now use such predicted images as input for existing place recognition algorithms.

TABLE II

OVERVIEW OF THE DEFAULT PARAMETERS

Parameter	Value
Number of Superpixels	1,000
Number of Words	10,000
Incorp. translations	all
Superpixel Descriptor	normalized Lab histogram (10 bins per channel), 128 Byte upright SURF descriptor, superpixel center y coordinate

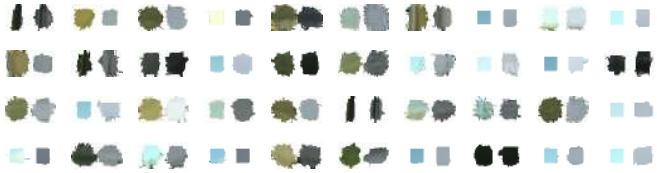

Fig. 8. Example words and their translations. Each tiny image pair shows a word from the spring vocabulary and the word from the winter vocabulary with the highest impact on the translation.

B. Experiments with FAB-MAP

In a first experiment we evaluated the performance of FAB-MAP [6] (using the openFAB-MAP implementation) on the Nordland dataset. We let FAB-MAP learn its visual vocabulary on either the spring training dataset, the winter training dataset or a combination of both.

As expected, directly matching winter against spring images was not successful: The maximum measured recall was 0.025 at 0.08 precision. This is presumably because FAB-MAP fails to detect common features in the images from both seasons.

The images produced by our proposed scene change prediction approach are not suitable for FAB-MAP since the patch structure of the synthesized images interferes with the necessary keypoint detection. In the following, we therefore examine two holistic approaches.

C. Extending and Improving BRIEF-Gist

BRIEF-Gist [21] is a so called holistic descriptor, i.e. a descriptor that describes the appearance of the complete image and not of single regions in it. The idea of a holistic scene descriptor was e.g. examined by Torralba et al. [24] with the introduction of the Gist descriptor. We chose the faster and more simple BRIEF-Gist descriptor on the opponency color space, using 32 bytes per channel.

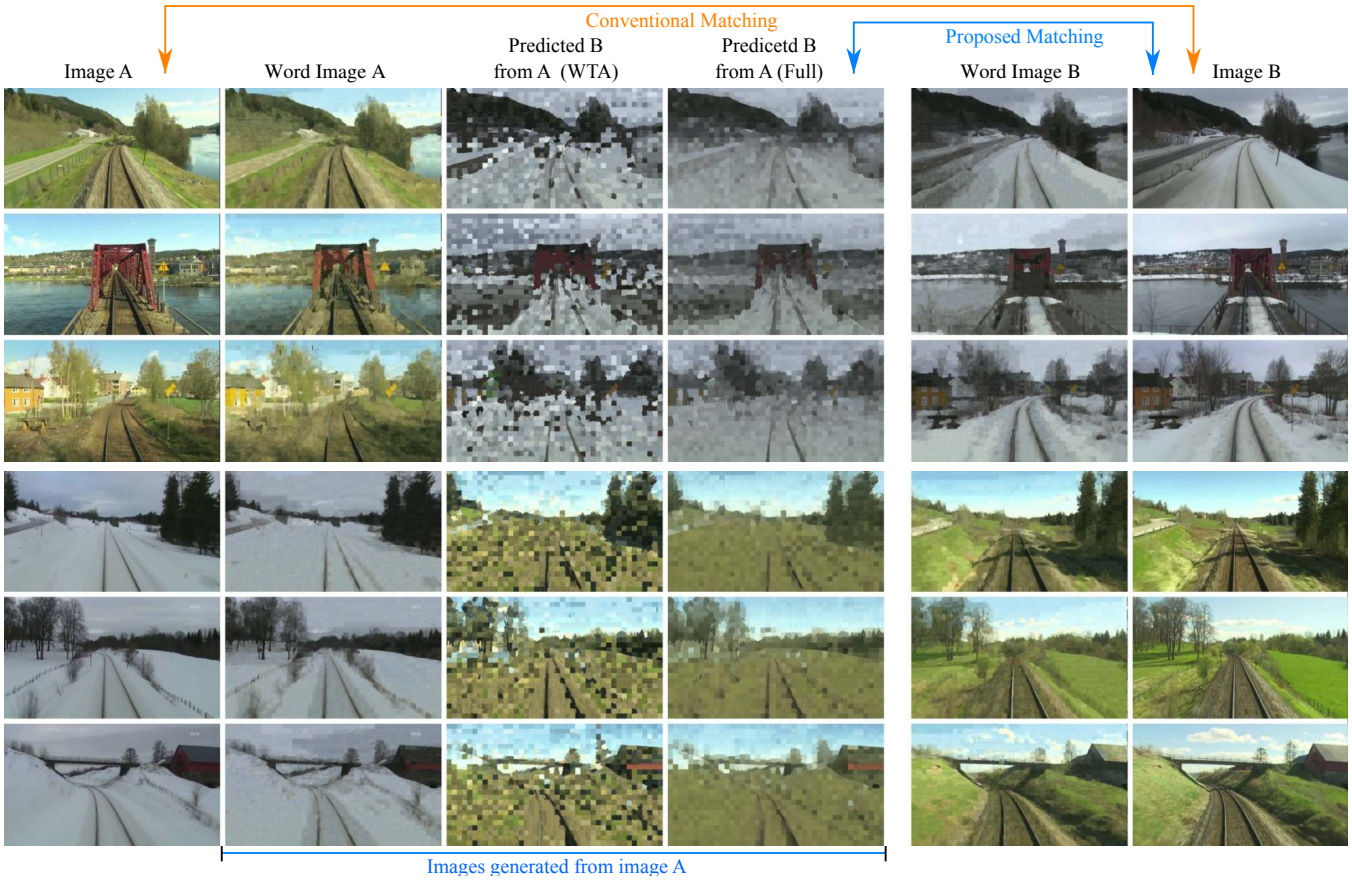

Fig. 9. Example images of the Nordland dataset, their word representations and predictions. The first column shows input query images A given to the prediction framework. The second column is a representation of the query image with words of the first vocabulary. All superpixel segments are replaced by word patches (word image). Applying a winner-takes-all dictionary (WTA) or a dictionary that uses the full distribution translates the words to the second vocabulary. Column three and four show the resulting predicted images B. For comparison column six shows the corresponding real image B and column five its word image representation. We propose not to match the visually very different images A and B directly, rather we propose to use a predicted image B for matching.

1) Experiments: In the following, the performance of BRIEF-Gist to recognize places of the Nordland dataset between spring and winter images is evaluated. We contrast the performance with and without the proposed prediction step and compare different setups of the prediction framework using the validation dataset. For each setup we compute a similarity matrix by comparing each combination of a spring and (potentially predicted) winter image. Since we know that spring and winter image sequences are synchronized, the ground truth similarity matrix is a diagonal matrix. For a quantitative evaluation we apply thresholds and compute precision-recall curves. Due to inaccuracies during synchronization and local self similarity we allow matchings of images with up to five frames distance in the sequence. To evaluate a setup of the prediction framework, we predict a winter image for each spring image based on the learned superpixel vocabularies and dictionary and use this for matching against the real winter images.

2) Results: The results of the evaluation with BRIEF-Gist are illustrated in Fig. 10. The red curve in Fig. 10 a) shows that due to the extreme appearance variations, direct matching of spring to winter images fails. However, the green curve shows the performance improvement when the proposed additional SP-ACP step is applied and matching is done between the winter and a *predicted* winter image. Although the recall at 100% precision does not benefit from the prediction, the maximum F-score improves from 0.14 to 0.31.

Fig. 10 b) compares the two proposed methods to build the dictionary, namely winner-takes-all (WTA) and storing the full probability distribution. The green curve in b) is the same as in a). From the red curve we can conclude that the WTA scheme has disadvantages in the important high precision area and storing the full distribution is beneficial.

Matching the predicted winter images against the *word* representation of the original winter images leads to a very similar loss of performance as can be seen in Fig. 10 c). To illustrate what is lost due to the transition from real images to word images, the blue curve in c) represents the performance of BRIEF-Gist when matching the spring images to their own (spring) word representation.

In a final experiment shown in d), we removed the y-coordinate from the superpixel descriptor. The red curve illustrates the slight performance drop if this additional

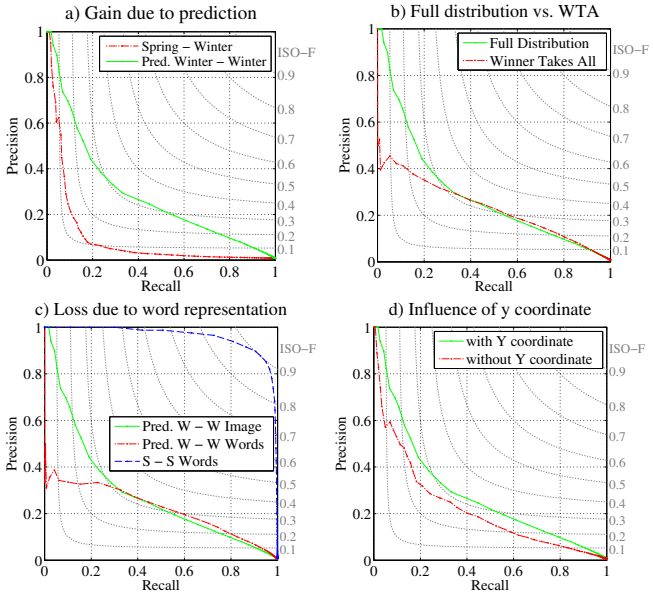

Fig. 10. Evaluation of the SP-ACP framework with BRIEF-Gist. a) Matching predicted winter images to winter images performs better than matching spring to winter images directly. b)-d) Comparison of several setups of the prediction framework. See text for details. Notice that the green curve represents the same setup in all plots.

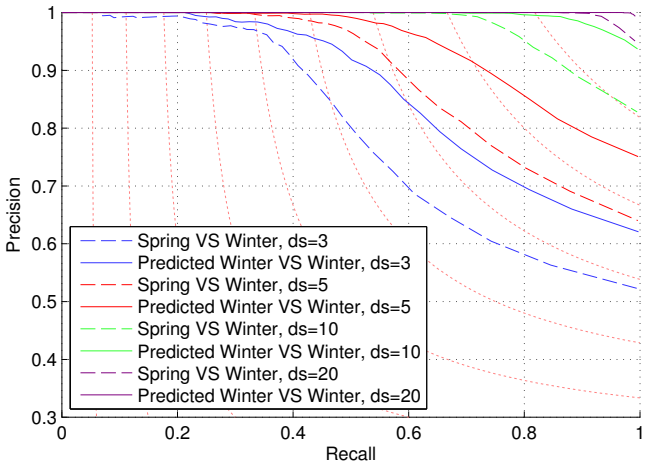

Fig. 11. Evaluation on the validation dataset. Precision recall plots obtained by combining SeqSLAM [17] with the proposed SP-ACP approach (solid lines) compared with SeqSLAM alone (dashed lines). Color indicates different trajectory lengths (d_s) used by SeqSLAM during the sequence matching. It is apparent that our proposed approach can significantly improve SeqSLAM's performance for all values of d_s. (For this experiment: #superpixels=2,500)

knowledge is omitted.

We can conclude that predicting the changed appearance of a scene improves the place recognition performance of BRIEF-Gist. This was clearly illustrated by Fig. 10 a). The best results were obtained when exploiting the full distribution over possible translations in the dictionary, matching predicted images against original images, and including the y-coordinate into the word descriptor.

D. Extending and Improving SeqSLAM

Published by Milford and Wyeth [17], SeqSLAM performs place recognition by matching whole *sequences* of images. This is in contrast to previous approaches like FAB-MAP or BRIEF-Gist that search for a *single* globally best match. [17] reported impressive recognition results on a dataset that contained footage recorded from a moving car during bright daylight and a rainy night in a suburban area. However, the matching performance comes at a price: SeqSLAM relies on relatively long sequences to be matched in order to reject false positive candidates. If loop closures in the trajectory form many but short overlapping sequences that are shorter than the required minimum length, SeqSLAM would fail. In order to be applicable in more general settings for long term navigation, this minimum sequence length has to be kept as short as possible.

Our goal is therefore to show that SeqSLAM's performance on short sequence lengths can be improved by combining it with our proposed scene change prediction.

1) Experimental Setup: SeqSLAM preprocesses the camera images by first downsampling them to e.g. 64×32 pixel before performing patch normalization. A simple sum of absolute differences measure determines the similarity between two images. Combining SeqSLAM with scene change prediction is particularly easy, since the change prediction algorithm can be executed as a preprocessing step before SeqSLAM starts with its own processing. Since in the experiments we attempted to match spring against winter images, we predicted the visual appearance of each spring scene in winter and fed the predicted winter images together with the original real winter images into SeqSLAM. We use the open source implementation OpenSeqSLAM [20] available online[3].

2) Results on the Validation Dataset: Fig. 11 compares the achieved results on the validation dataset similar to the experiments using BRIEF-Gist. The precision-recall plot shows the performance of SeqSLAM alone (i.e. *without* scene change prediction) using the dashed lines. The precision-recall curves for the combination of SP-ACP and SeqSLAM are drawn with solid lines. We show the results for different settings of the SeqSLAM's trajectory length parameter d_s, as indicated by the different colors.

The apparent result is that SeqSLAM can immediately benefit from the change prediction. The gain in precision and recall, as well as the increased recall at 100% precision is visible for all trajectory lengths d_s. The F-score increases by almost 0.1 for short and mid sequence lengths and tends towards 1 for the longest length. Notice that $d_s = 20$ corresponds to a trajectory length of 10 seconds, since the validation data was captured with 2 Hz from the original video footage.

We have to remark that the Nordland dataset is perfectly suited for SeqSLAM since the whole dataset consists of one single long sequence and the camera observes the scene from

[3] https://openslam.org/openseqslam.html

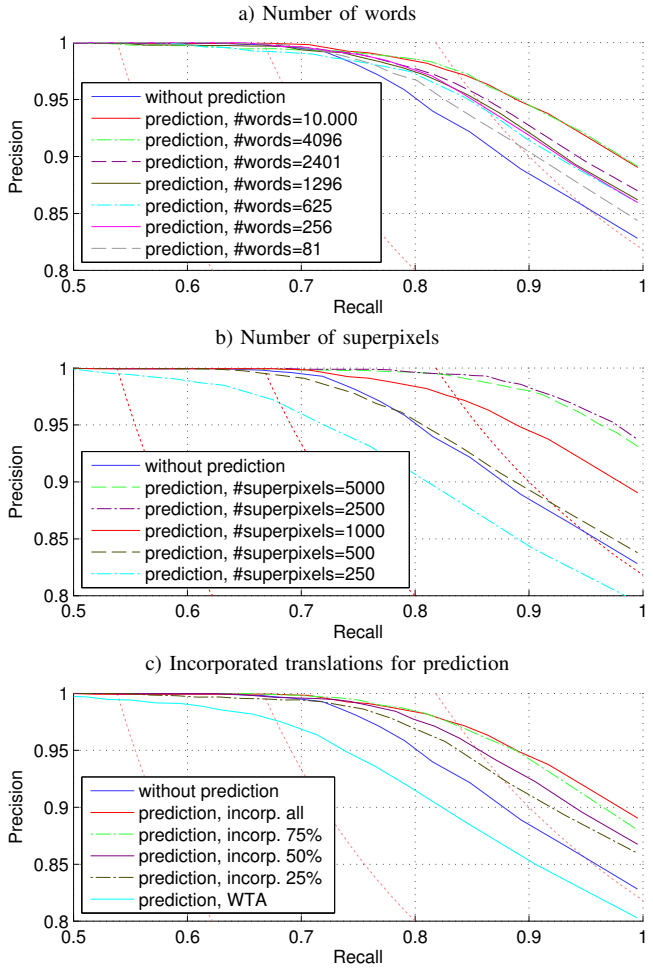

Fig. 12. Parameter evaluation of the proposed prediction framework using SeqSLAM on the Nordland validation dataset. Starting from a default set of parameters, we vary a single parameter to investigate its influence on the overall performance (see text for details). In each plot, the blue solid line is the performance without the proposed prediction framework and the solid red line shows the same default parameter setup of the prediction framework in all plots (#words=10,000, #superpixels=1,000, incorporate all translations during prediction).

almost exactly the same viewpoint in all four seasons as the train follows its tracks. Even more, the velocity of the train is equal most of the time in all seasons. In robotic applications these conditions would usually not be met and we can expect SeqSLAM in its current form to perform worse in general. Our point however, was to show that SeqSLAM can in any case benefit from a combination with the proposed appearance change prediction.

We can conclude that although SeqSLAM alone reaches good matching results, they can be significantly improved by first predicting the appearance of the query scene under the viewing conditions of the stored database scenes.

In the following, we are going to evaluate various influences on the prediction quality and the resulting place recognition performance using the validation dataset. This is followed by final results on the complete Nordland track using the test dataset.

3) Parameter Evaluation on the Validation Dataset: Besides the characteristics of the training data, the number of words in the learned vocabularies, the number of superpixel segments per image, and the amount of incorporated transitions in the prediction are important parameters of the proposed SP-ACP system. Starting from a default parameter setting (#words=10,000, #superpixels=1,000, incorporate all translations during prediction) we vary each of these three parameters to evaluate its influence on the prediction and resulting place recognition performance. Fig. 12 shows results of the conducted experiments based on the spring and winter training and validation datasets.

a) Number of visual words: To evaluate the influence of the number of words in the visual vocabularies, we varied the branching factor of the hierarchical k-means we used to cluster the superpixel descriptors. Since the depth of the tree was held constant (depth 4), this resulted in 81 to 10,000 words. Fig. 12 *a)* shows an obvious trend that more words perform better. However, there are some peculiarities: e.g. 256 words perform better than 625, and the performance for 4096 and 10,000 words is almost identical. We assume this results from the limited amount of training data in our experiments. Supposedly we cannot expect to learn the true full distribution to translate between two 10,000 word vocabularies from only one million training samples.

b) Number of superpixels: Fig. 12 *b)* shows the influence of the number of superpixel segments per image. We can observe an expected trade-off in the performance: The higher the number of superpixels, the better are object boundaries covered. However, this also means smaller superpixels, that cover less image content and are less meaningful. In our experiments we observed that the optimal number of superpixels is 2500 per image.

c) Incorporated translations: Fig. 12 *c)* evaluates the last of the three parameters: the percentage of words incorporated in the prediction.

Remember from section III-C that in order to predict an image from e.g. summer conditions to winter conditions, we first compute a word representation of the summer image and then synthesize a winter representation for each of the summer words using the learned dictionary. During this synthesis, we can either use only the single winter word that translates best according to the training data (winner takes all, WTA); or we can use a weighted combination of the words that explain e.g. 50% of the transitions from the training data or even a weighted combination of all words. The results of this comparison is illustrated in 12 *c)*. The obvious conclusion is that incorporating more words yields better results, however, incorporating more than 75% of the probability mass does not yield much improvement.

In this setup, WTA breaks the prediction. The similar experiment using BRIEF-Gist (see Fig. 10 *b)*) shows also problems in the high precision regime but a clear benefit in the mid- and high-recall regimes. The example prediction results using WTA and the full distribution in Fig. 9 show that the WTA predictions have higher contrast between neighbored patches while the predictions from the full dis-

tribution are more smoothed. These high local contrasts may have a negative influence in combination with the local patch normalization of SeqSLAM.

d) Dataset characteristics: Fig. 13 shows another important factor for the influence of the proposed prediction step for place recognition: the characteristics of the dataset. These figures show the results for applying SP-ACP and SeqSLAM on other combinations of seasons (i.e. other than spring-winter). Our previous work [20] explored the performance of SeqSLAM (*without* SP-ACP) for the various seasonal combinations and found that fall-winter was the most difficult and summer-fall was the easiest combination for place recognition using SeqSLAM without appearance change prediction. We therefore chose these two combinations for comparison.

Both plots in Fig. 13 illustrate the results for various sequence lengths with and without the prediction step. We can see that independent of using the prediction or not, finding correct matchings between summer and fall is much easier than between fall and winter. This is an expected result. However, although using the additional prediction step improves performance for the difficult (fall-winter) case, the performance actually decreases for the easy (summer-fall) case. Although the overall performance still remains reasonable, this result needs to be explained: A look at example images in Fig. 14 shows that the predicted fall images are visually very similar to the real fall images. However, we suppose that the smoothing and artifacts introduced by the SP-ACP prediction step are a drawback in comparison to the high similarity between original summer and fall images. The same effect is also visible on the comparison of spring images and their word representation using BRIEF-Gist in Fig. 10 *c)*. Although these disturbing factors are the same for the comparison between fall and winter images, the prediction still introduces a visually evident benefit since the original imaging conditions are very different.

In the conducted place recognition experiments the prediction step improves the place recognition performance with a prediction of winter from fall but the benefit is not as large as for predicting winter from spring. This is due to the higher diversity in appearance of trees, bushes, meadow etc. in the spring images compared to the fall images. The example images of spring and fall in figures 9 and 14 give an impression of the richer information in the spring images. This enables the proposed prediction system to better learn the different appearance changes of different image content and thus to produce better predictions from the spring images compared to the predictions from fall images. Incorporating more contextual information (e.g surrounding image patches or high level knowledge) could help to better exploit the provided appearance diversity. Directions for future work in this direction can be found in section VI.

4) Final Results on the 728 km Test Dataset: The final result of this work is a place recognition experiment using SeqSLAM on the complete 728 km Nordland track between spring and winter using the *test* dataset of section IV-B. The training dataset remains the same as for the previous experiments. The ratio of training and test dataset is illustrated in Fig. 7, where the red part indicates the training part and the test dataset is shown blue. Fig. 15 shows the result. Again we compare the place recognition performance with (solid) and without (dashed) the prediction for different values for the SeqSLAM sequence length ds. Since the test dataset is recorded with 0.1 frames per second, a sequence length of 3 corresponds to a 30 seconds trajectory in the original video.

It is apparent how SeqSLAM benefits from the proposed prediction step for all sequence lengths on the test dataset. Both the maximum F-score and the recall at 100% pre-

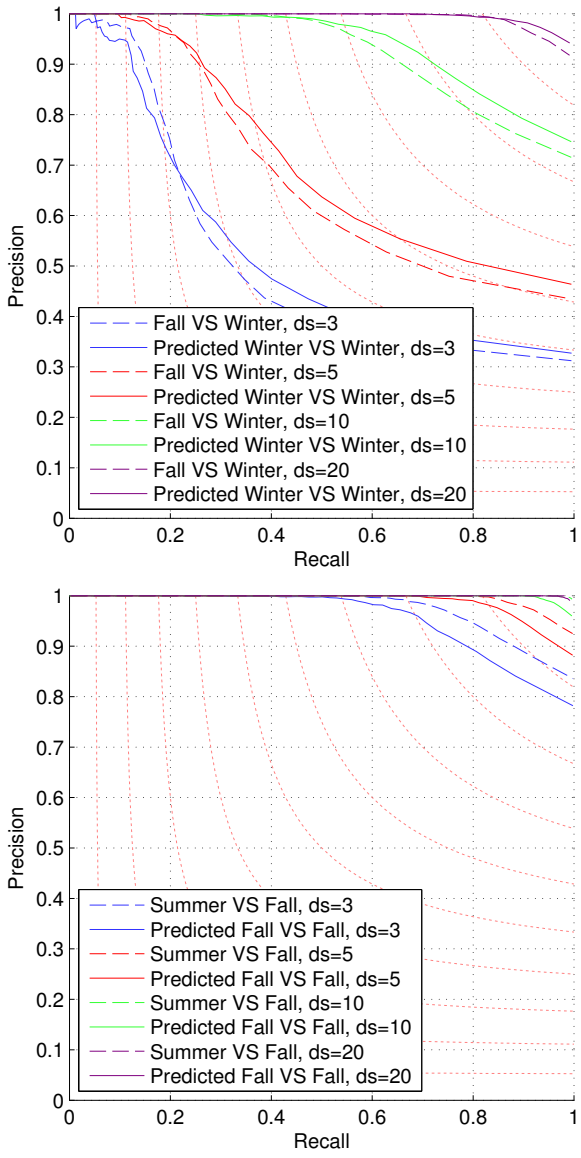

Fig. 13. Predicting between different seasons. Evaluation based on SeqSLAM and the validation dataset. *(left)* Place recognition between fall and winter is a challenging task. Here the prediction framework can help to improve the overall performance. *(right)* In contrast, places in summer and fall (at least in the Norland dataset) are more similar and place recognition is much easier (see Fig. 14). In such cases, the prediction result can become worse. However, the overall place recognition performance remains reasonable. (For these experiments: #superpixels=2,500)

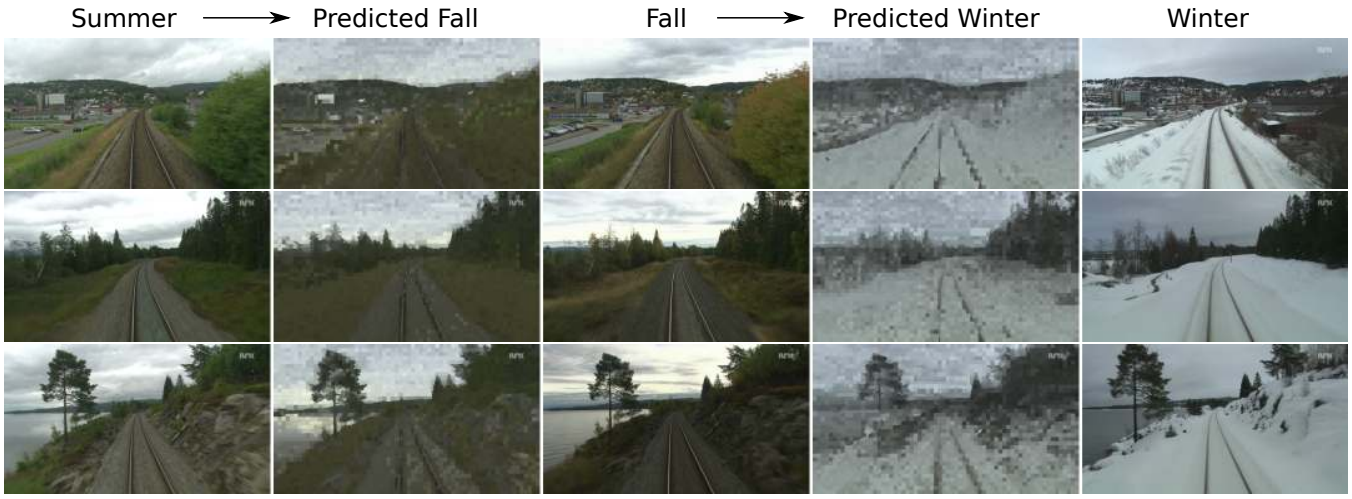

Fig. 14. Example images and predictions for summer, fall and winter. Each row shows the same scene at three different seasons. The images between the original Summer, Fall and Winter images are predictions from summer to fall and from fall to winter.

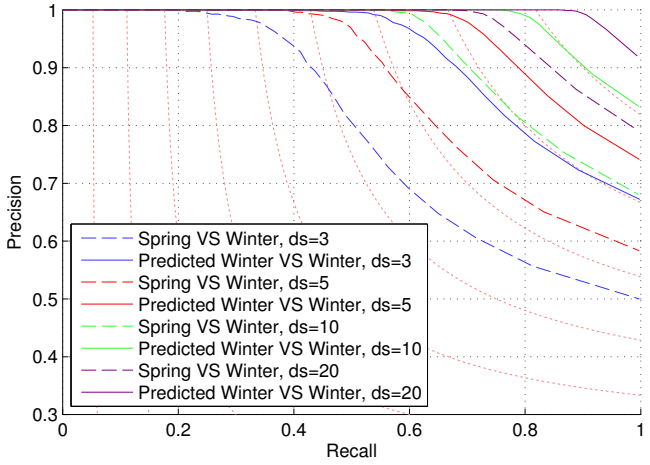

Fig. 15. Final result on the complete Nordland track (*test dataset*). Again, we compare the performance of the combination of SeqSLAM [17] with the proposed scene change prediction approach (solid lines) and SeqSLAM alone (dashed lines). Color indicates different trajectory lengths (d_s) used by SeqSLAM during the sequence matching. The proposed prediction framework, trained on an 8 minutes subset, can significantly improve the place recognition performance on the complete track. (For this experiment: #superpixels=2,500)

cision improve significantly. This is a remarkable result, since it shows that our proposed SP-ACP system is able to extract enough knowledge from the training dataset of only 8 minutes to significantly improve the place recognition on the complete journey of almost 10 hours. Considering the very different environmental conditions met along the journey (e.g. lowlands, highlands, mountains) and the rather homogeneous appearance (lowlands) in the training dataset makes this result even more remarkable. However, we have to clearly point out that this generalization capability is of course limited to environmental conditions with similar systematic change. The proposed system can not be learned on the Nordland dataset and applied to images of e.g. Manhattan.

VI. Current Limitations of the Approach and Future Work

The proposed SP-ACP system is a rather straightforward implementation of the idea of incorporating an additional prediction step for place recognition in systematically changing environments. However, there is plenty of space for improvements.

Obviously the prediction step incorporates smoothing and artifacts in the predicted images. This can cause a decrease in place recognition performance if the compared original sequences are very similar (e.g. summer and fall). However, the predicted images are visually appealing and the place recognition performance remains reasonable. This can be interpreted as a kind of "upper bound" for the recognition performance which is introduced by the smoothing and the artifacts of the prediction.

In its current form, our algorithm requires perfectly (near pixel-accurate) aligned images in the training phase. This requirement is clearly a key limitation of the proposed approach, since it is hard to fulfill and limits the available training datasets. We will explore ways to ease or overcome this requirement in future work, e.g. by anchoring the training images on stable features. This would increase the availability of potential training datasets collected by robots or vehicles in realistic scenarios that are close to real-world applications.

Currently, we synthesize an actual image during the prediction. This simplifies the qualitative evaluation by visually comparing the predicted with the real images and further allows to use existing place recognition algorithms for quantitative evaluation. However, the proposed idea of scene change prediction can in general be performed on different levels of abstraction: It could also be applied directly on holistic descriptors like BRIEF-Gist, on visual words like the ones used by FAB-MAP or on the downsampled and patch-normalized thumbnail images used by SeqSLAM.

Furthermore, the learned dictionary can be as simple as a

one-to-one association (like the mentioned winner-takes-all scheme) or capture a full distribution over possible translations for a specific word. In future work this distribution could also be conditioned on the state of neighboring segments, and other local and global image features and thereby incorporate mutual influences and semantic knowledge. This could be interpreted as introducing a *grammar* in addition to the vocabularies and dictionaries. How such extended statistics can be learned from training data efficiently is an interesting direction for future work.

If the dictionary does not exploit such higher level knowledge (as in the superpixel implementation introduced here) the quality of the prediction is limited. In particular, when solely relying on local appearance of image segments for prediction, the choice of the training data is crucial. It is especially important that the training set is from the same domain as the desired application, since image modalities that were not well-covered by the training data can not be correctly modeled and predicted.

Exploring the requirements for the training dataset and how the learned vocabularies and dictionary can best generalize between different environments will be an important part of our future research. A first step into analysing how well the system can generalize is to train and test it on a more diverse dataset. We therefore collected imagery from webcams around the world for several months that comprises different seasons, weather conditions and times of day.

A further limitation of the system in its current form is that it requires different vocabularies for *discrete* sets of environmental conditions. While it is of course possible to create and manage a larger number of such vocabularies and the respective mutual dictionaries, a unified approach would be more desirable.

As already mentioned, the Nordland dataset provides somewhat optimal conditions (apart from the season-induced appearance changes) for place recognitions, since the camera observes the scene from almost exactly the same viewpoint in all four seasons and the variability of the scenes in terms of semantic categories is rather low. These conditions would usually not be met in a typical robotic application and we therefore prepare to evaluate the proposed approach in a more general setting using data from real robots in different environments, and vehicles in urban settings.

VII. Conclusions

Our paper described the novel concept of learning to predict systematic changes in the appearance of environments. We explained our SP-ACP implementation based on superpixel vocabularies and provided examples for scene change prediction between different seasons. We furthermore demonstrated how two approaches to place recognition, BRIEF-Gist and SeqSLAM, can benefit from the scene change prediction step.

We evaluated all important parameters of the proposed system and found that none of them requires particular careful tuning. Parameter values can be chosen safely from a wide range of values, and the system still produces useful predictions for place recognition. To conclude: using many words (e.g. 10,000), many superpixels (e.g. 2,500) and incorporating many translations (e.g. 75%) leads to good prediction results. These insights helped to demonstrate new results with a significant improvement of the place recognition performance of SeqSLAM on the complete Nordland track. Predictions from spring to winter learned on a small 8 minute subset of the available data yield better matchings on the complete 728 km track using the same sequence length d_s or allow to use shorter sequences for the same recognition performance.

References

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] Hernán Badino, Daniel F. Huber, and Takeo Kanade. Real-time topometric localization. In *International Conference on Robotics and Automation (ICRA)*, 2012.

[3] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, March 2003.

[4] Peter Biber and Tom Duckett. Experimental Analysis of Sample-Based Maps for Long-Term SLAM. *International Journal of Robotics Research*, 28(1):20–33, January 2009.

[5] Winston Churchill and Paul M. Newman. Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation. In *International Conference on Robotics and Automation (ICRA)*, 2012.

[6] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *International Journal of Robotics Research*, 27(6):647–665, June 2008.

[7] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. In *28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH, pages 341–346. ACM, 2001.

[8] Arren Glover, William Maddern, Michael Milford, and Gordon Wyeth. FAB-MAP + RatSLAM : appearance-based SLAM for multiple times of day. In *Int. Conf. on Robotics and Automation (ICRA)*, 2010.

[9] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-Class Segmentation with Relative Location Prior. *International Journal of Computer Vision*, 80(3):300–316, December 2008.

[10] Xuming He, Richard S. Zemel, and Volodymyr Mnih. Topological map learning from outdoor image sequences. *Journal of Field Robotics*, 23(11-12):1091–1104, 2006.

[11] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. In *28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH, pages 327–340. ACM, 2001.

[12] H. Johannsson, M. Kaess, M.F. Fallon, and J.J. Leonard. Temporally scalable visual SLAM using a reduced pose graph. In *RSS Workshop on Long-term Operation of Auton. Robotic Systems in Changing Environments*, 2012.

[13] Kurt Konolige and James Bowman. Towards lifelong visual maps. In *International Conference on Intelligent Robots and Systems (IROS)*, 2009.

[14] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T. Freeman. SIFT Flow: Dense Correspondence across Different Scenes. Europ. Conf. on Comp. Vision (ECCV), 2008.

[15] William Maddern and Stephen Vidas. Towards robust night and day place recognition using visible and thermal imaging. In *Robotics Science and Systems Conference (RSS)*, 2012.

[16] Michael Milford and Gordon Wyeth. Persistent Navigation and Mapping using a Biologically Inspired SLAM System. *International Journal of Robotics Research*, 29(9):1131–1153, August 2010.

[17] Michael Milford and Gordon Fraser Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2012.

[18] Peer Neubert, Niko Sünderhauf, and Peter Protzel. Appearance Change Prediction for Long-Term Navigation Across Seasons. In *European Conference on Mobile Robotics (ECMR)*, 2013.

[19] Xiaofeng Ren and Jitendra Malik. Learning a Classification Model for Segmentation. International Conference on Computer Vision (ICCV), 2003.

[20] Niko Sünderhauf, Peer Neubert, and Peter 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.

[21] Niko Sünderhauf and Peter Protzel. BRIEF-Gist - Closing the loop by simple means. In *International Conference on Intelligent Robots and Systems (IROS)*, 2011.

[22] Niko Sünderhauf and Peter Protzel. Switchable Constraints vs. Max-Mixture Models vs. RRR – A Comparison of three Approaches to Robust Pose Graph SLAM. In *Proc. of Intl. Conf. on Robotics and Automation (ICRA)*, 2013.

[23] Joseph Tighe and Svetlana Lazebnik. Superparsing: scalable non-parametric image parsing with superpixels. European Conference on Computer Vision (ECCV), 2010.

[24] Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin. Context-based vision system for place and object recognition. International Conference on Computer Vision (ICCV), 2003.

[25] Christoffer Valgren and Achim J. Lilienthal. SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments. *Robotics and Autonomous Systems*, 58(2):149–156, February 2010.

[26] Wei Zhang and Jana Kosecka. Localization Based on Building Recognition. Conf. on Comp. Vision a. Pattern Recognition (CVPR), 2005.