To appear in Proc. of European Conference on Mobile Robotics (ECMR), 2013. DOI: not yet available


# Appearance Change Prediction for Long-Term Navigation Across Seasons

Peer Neubert, Niko Sünderhauf and Peter Protzel

***Abstract*— Changing environments pose a serious problem to current robotic systems aiming at long term operation. While place recognition systems perform reasonably well in static or low-dynamic environments, severe appearance changes that occur between day and night, between different seasons or different local weather conditions remain a challenge. In this paper we propose to learn to *predict* the changes in an environment. Our key insight is that the occurring appearance 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 a novel implementation based on vocabularies of superpixels (SP-ACP). Despite its simplicity, we can further show that the proposed approach can improve the performance of SeqSLAM and BRIEF-Gist for place recognition on a large-scale dataset that traverses an environment under extremely different conditions in winter and summer.**

## 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 [14] 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, 09126 Chemnitz, Germany. peer.neubert@etit.tu-chemnitz.de

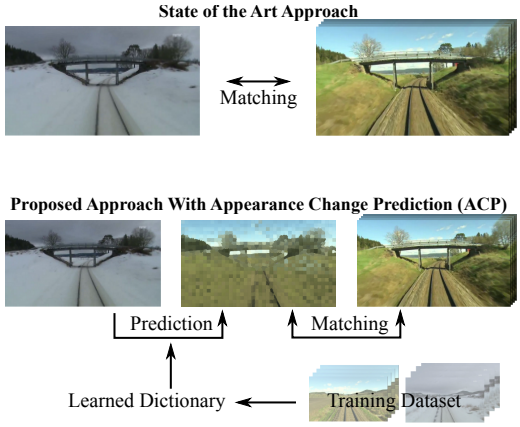


Fig. 1. State of the art approaches to place recognition will attempt to directly match two scenes even if they have been observed under extremely different environmental conditions. This is prone to error and leads to bad recognition results. Instead, we propose to incorporate an intermediate appearance change prediction (ACP) step 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 learned dictionary that exploits the systematic nature of the seasonal changes.

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 incorporate an intermediate appearance change prediction (ACP) step 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.

To the best of our knowledge, the idea of predicting extreme appearance changes across seasons to aid place recognition is novel and has not been proposed before. It is the main contribution of this paper. Furthermore, we prove the feasibility of our idea and describe an implementation based on superpixel vocabularies (SP-ACP). 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 [14] and BRIEF-Gist [15] on a new, publicly available large-scale dataset that traverses an environment in winter and summer.

## II. Related Work

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. Based on local keypoint features, Valgren and Lilienthal [18] 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. Glover et al. [1] present a combination of the advanced local feature recognition system FAB-MAP [6] and the biologically inspired SLAM approach RatSLAM [13] 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 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. The pose cell filtering of RatSLAM is a step towards using sequences for matching. In their subsequent work, the RatSLAM authors presented SeqSLAM [14] 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. [4] 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. Maddern and Vidas [12] 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. He et al. [19] 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. 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, 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 [10], or views [11] have to be introduced to the map can be reduced.

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* [9]: Given images $(A_S, A_W)$ 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 appearance change prediction 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. [9] create visually appealing results but have not yet been used in context of place recognition. They focus on using

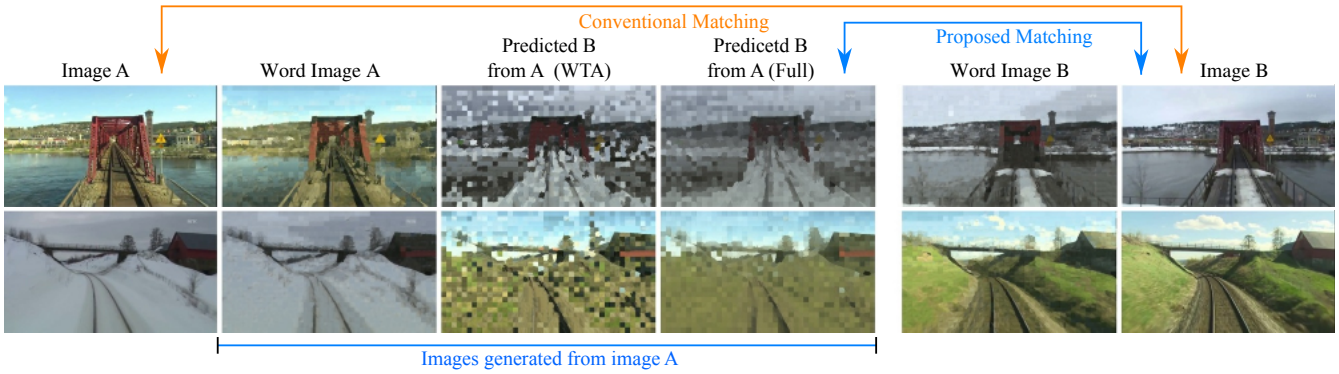


Fig. 2. Example images of the Nordland dataset, their word representations and predictions. The first column shows input query images A given to the SP-ACP framework. The second column is a representation of the query image with words of the first vocabulary. All superpixel segments are replaced by word 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.

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 Appearance Changes Across Seasons

In this section we explore how the changing appearance of a scene across different environmental conditions can be predicted. Throughout the remainder of the paper 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 as image parts and cluster them to vocabularies using a descriptor. 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.

Fig. 3 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 segmented the images into 1000 superpixels using SLIC [2] and learned 10.000 words for each vocabulary. 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. [17] combine shape, location, texture (using SIFT) and color features. Barnard et al. [3] use 40 features, the descriptor of Gould et al. [8] even includes multiple color descriptors. 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. 5. We do not apply further dimensionality reduction.

### 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 if Fig. 3. Since the images from

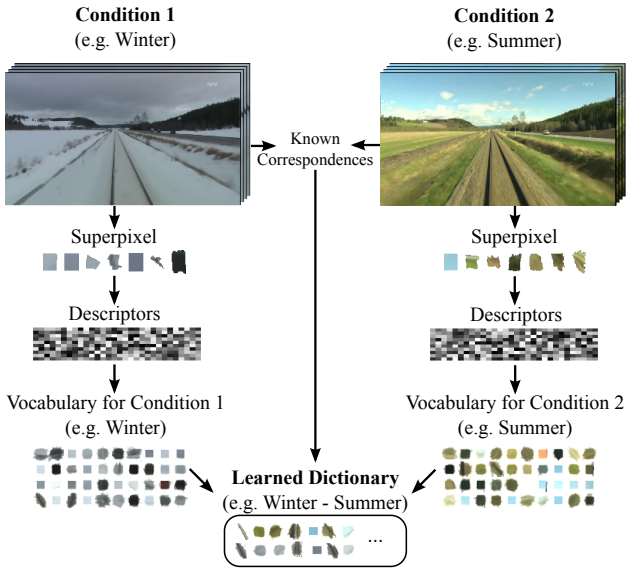


Fig. 3. SP-ACP learning a dictionary between images under different environmental conditions (e.g. winter and summer). The images are first segmented into superpixels and a descriptor is calculated for each superpixel. These descriptors are then clustered to obtain a vocabulary of visual words for each 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.

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 IV will compare both approaches.

### C. Predicting Image Appearances Across Seasons

Fig. 4 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. 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

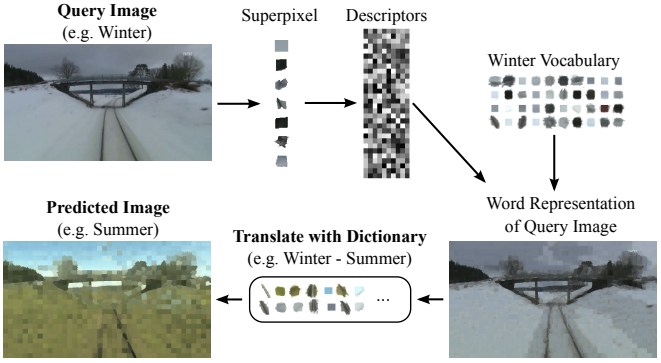


Fig. 4. SP-ACP predicting the appearance of a query image under different environmental conditions: How would the current winter scene appear in summer? The query image is first segmented into superpixels and a descriptor is calculated for each of these segments. With this descriptor each superpixel 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. 2). The result of the process is a synthesized image that predicts the appearance of the winter query image in summer time.

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, 9]) is done to improve the appearance or smoothness of the resulting word images. Example word images and predictions are shown in Fig. 2.

## IV. Experiments and Results

### A. The Nordland Dataset

The TV documentary "Norlandsbanen – Minutt for Minutt" by the Norwegian Broadcasting Corporation NRK provides video footage of the 728 km long train ride between the Norwegian cities of Trondheim and Bodø[1]. The complete 10 hour journey has been filmed from the perspective of the train driver four times in spring, summer, fall, and winter. The full-HD recordings have been time-synchronized and contain GPS data. To form the training dataset, we extracted approximately 900 frames from the first 8 minutes of a 30 minutes subset. This training dataset was used to learn the visual vocabulary for summer and winter and the dictionary to translate between both seasons. The remaining 22 minutes of the video subset served as the test dataset.

### B. Experiments with FAB-MAP

In a first experiment we evaluated the performance of FAB-MAP [6] (using the openFAB-MAP implementation) on the dataset. We let FAB-MAP learn its visual vocabulary on either the summer training dataset, the winter training dataset or a combination of both. As expected, directly matching winter against summer 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 SP-ACP approach are not suitable

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

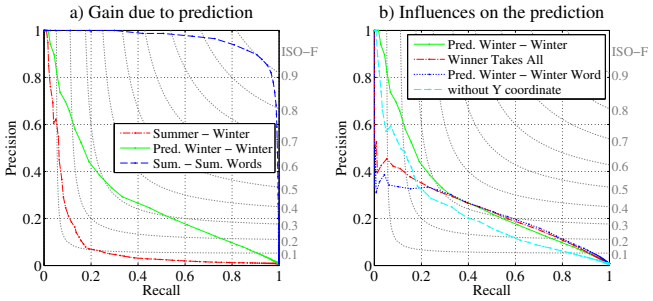


Fig. 5. Evaluation of the SP-ACP framework with BRIEF-Gist. a) Matching predicted winter images to winter images performs better than matching summer to winter images directly. b) Comparison of several setups of SP-ACP framework. See text for details. Notice that the green curve represents the same setup in both plots.

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

In the following, the performance of BRIEF-Gist [15] to recognize places of the Nordland dataset between summer and winter images is evaluated. We compute BRIEF-Gist on the opponency color space, using 32 bytes per channel. We contrast the performance with and without the proposed prediction step and compare different setups of the prediction framework. For each setup we compute a similarity matrix by comparing each combination of a summer and (potentially predicted) winter image. Since we know that summer and winter image sequences are synchronized, the ground truth similarity matrix is a diagonal matrix. For 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 summer image based on the learned superpixel vocabularies and dictionary and use this for matching against the real winter images.

The results of the evaluation with BRIEF-Gist are illustrated in Fig. 5. The red curve in Fig. 5 a) shows that due to the extreme appearance variations, direct matching of summer 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. Both recall and precision improve. To illustrate what is lost due to the transition from real images to word images, the blue curve in a) represents the performance of BRIEF-Gist when matching the summer images to their own (summer) word representation. Fig. 5 b) evaluates several influcences on the quality of the prediction. From the red curve we can conclude that the Winner Takes All scheme has disadvantages in the high precision area and storing the full distribution (as for the green curve) 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 the blue curve. In a final experiment we removed the $y$-coordinate from the superpixel descriptor. The cyan curve illustrates the slight performance drop if this additional knowledge is omitted. We can conclude that predicting the changed appearance of a scene significantly improves the place recognition performance of BRIEF-Gist. This was clearly illustrated by Fig. 5 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 [14], 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. [14] 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 appearance change prediction.

*1) Experiments:* SeqSLAM preprocesses the camera images by first downsampling them to e.g. $64 \times 32$ pixel before performing patch normalization. A simple sum of absolute differences measure determines the similarity between two images. Combining SeqSLAM with SP-ACP 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 summer against winter images, we predicted the visual appearance of each summer scene in winter and fed the predicted winter images together with the original real winter images into SeqSLAM.

*2) Results:* Fig. 6 compares the achieved results obtained using the OpenSeqSLAM implementation [16]. The precision-recall plot shows the performance of SeqSLAM alone (i.e. *without* appearance change prediction) using the solid lines. The precision-recall curves for the combination of SP-ACP and SeqSLAM are drawn with dashed 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 is visible for all trajectory lengths $d_s$. It is most noticeable for the red curve ($d_s = 20$) where the f-score increases by more than $0.05$. Notice that $d_s = 20$ corresponds to a trajectory length of 10 seconds, since the test data was captured with 2 Hz from the original video footage.

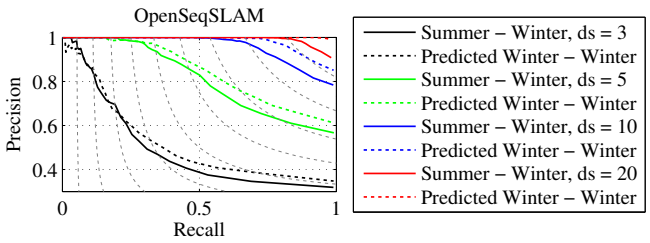


Fig. 6. Precision recall plots obtained by combining SeqSLAM [14] with the proposed SC-ACP (dashed lines) compared with SeqSLAM alone (solid lines). Color indicates different trajectory lengths ($d_s$) used by SeqSLAM during the sequence matching. It is apparent that our proposed approach can significantly improve SeqSLAM's performance at all sequence lengths.

We have to remark that the Nordland dataset is perfectly suited for SeqSLAM since the whole dataset consists of one single long sequence, the images of all runs are time synchronized and the camera observes the scene from almost exactly the same viewpoint in all four seasons as the train follows its tracks. 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.

## V. Discussion and Conclusions

Our paper described the novel concept of appearance change prediction (ACP): learning to predict systematic changes in the appearance of environments. We explained a direct implementation of the idea based on superpixel vocabularies (SP-ACP). We furthermore demonstrated how two approaches to place recognition, BRIEF-Gist and SeqSLAM, can benefit from the appearance change prediction step. Using SP-ACP, we can 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 ACP 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. 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. Furthermore, currently the training requires perfectly aligned image pairs. Exploring the requirements for the training dataset and how the learned vocabularies and dictionary can best generalize between different environments will be part of our future research. A key 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.

## References


[1] FAB-MAP + RatSLAM : appearance-based SLAM for multiple times of day, author = A. Glover and W. Maddern and M. Milford and G. Wyeth, year = 2010,. In *Int. Conf. on Rob. a. Autom. (ICRA)*.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. on Pat. Anal. and Mach. Intel.*, 34, 2012.

[3] Kobus B., P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3, 2003.

[4] H. Badino, D. Huber, and T. Kanade. Real-time topometric localization. In *Int. Conf. on Rob. a. Autom. (ICRA)*, 2012.

[5] 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.

[6] M. Cummins and P. Newman. Fab-map.

[7] A. Efros and W. Freeman. Image quilting for texture synthesis and transfer. In *SIGGRAPH*, 2001.

[8] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-Class Segmentation with Relative Location Prior. *Int. J. Comput. Vision*, 80, 2008.

[9] A. Hertzmann, C. Jacobs, N. Oliver, B. Curless, and D. Salesin. Image analogies. In *SIGGRAPH*, 2001.

[10] 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.

[11] K. Konolige and J. Bowman. Towards lifelong visual maps. In *Int. Conf. on Intel. Robots and Systems (IROS)*, 2009.

[12] W. Maddern and S. Vidas. Towards robust night and day place recognition using visible and thermal imaging. In *Robotics Science a. Syst. Conf. (RSS)*, 2012.

[13] M. Milford and G. Wyeth. Persistent navigation and mapping using a biologically inspired slam system. *Int. J. Rob. Res.*, 2010.

[14] 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)*. IEEE, 2012.

[15] N. Sünderhauf and P. Protzel. BRIEF-Gist - Closing the loop by simple means. In *Int. Conf. on Intel. Rob. and Syst. (IROS)*, 2011.

[16] N. Sünderhauf, P. Neubert, and P. Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. In *Proc. of Workshop on Long-Term Autonomy at Int. Conf. on Rob. a. Autom. (ICRA)*, 2013.

[17] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. Europ. Conf. on Comp. Vision (ECCV), 2010.

[18] C. Valgren and A. Lilienthal. SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments. *Robot. Auton. Syst.*, (2), 2010.

[19] X.He, R. Zemel, and V. Mnih. Topological map learning from outdoor image sequences. *J. Field Robotics*, 23, 2006.