Predicting the Change – A Step Towards Life-Long Operation in Everyday Environments

Niko Sünderhauf, Peer Neubert, Peter Protzel
Department of Electrical Engineering and Information Technology
Chemnitz University of Technology, Germany
{niko.suenderhauf, peer.neubert, peter.protzel}@etit.tu-chemnitz.de

***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 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 novel idea of scene change prediction and a proof of concept implementation based on vocabularies of superpixels. We can show that the proposed approach improves 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 operation in changing environments is one of the major challenges in robotics today. Robots operating autonomously over the course of days, weeks, and months are faced 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 [3] 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 [2] 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.

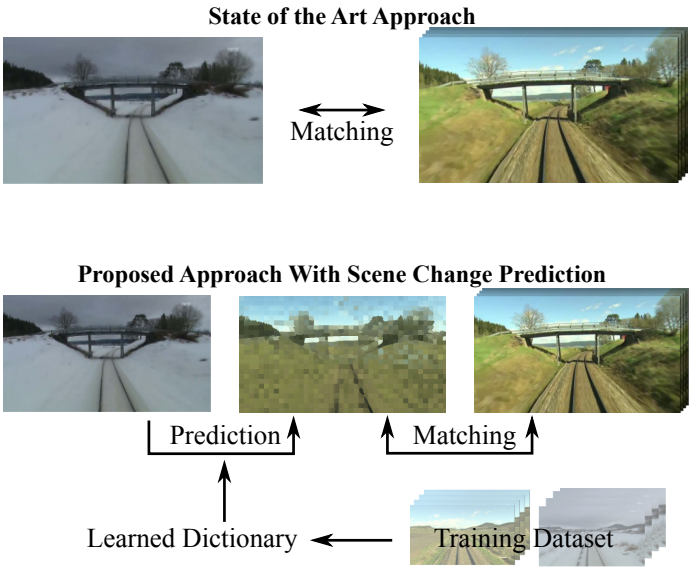

Fig. 1. State of the art approaches to place recognition attempt to directly match two scenes, even if they have been observed under extremely different environmental conditions. This is prone to error and leads to bad recognition results. Instead, we propose to *predict* how the query scene (the winter 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 the paper 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* or *hallucinate* 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 depends on the actual place recognition algorithm that is applied. When using approaches like SeqSLAM [3] or BRIEF-Gist [4], the result would be a synthesized *image* as illustrated in Fig. 1. This image preserves the structure of the original scene but is close in visual appearance to the corresponding original summer scene. When using place recognition based on a bag of words approach (e.g. FAB-MAP), the result of the prediction process would be a translated bag of words.

In any case, the proposed 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 scene changes across seasons to aid place recognition is novel and has not been proposed before.

II. Learning to Predict Scene Changes Across Seasons

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. However, 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. This idea can be extended to groups of parts, incorporating their mutual relationships.

To predict how the appearance of a scene changes between different conditions (e.g. summer and winter), we propose to first conduct a learning phase on training data. This data comprises scenes observed under both summer and winter conditions. In the subsequent prediction phase, the change in appearance of a new scene can be predicted using the results of the training phase. In the following, we explain our current proof of concept implementation of the proposed scene change prediction approach.

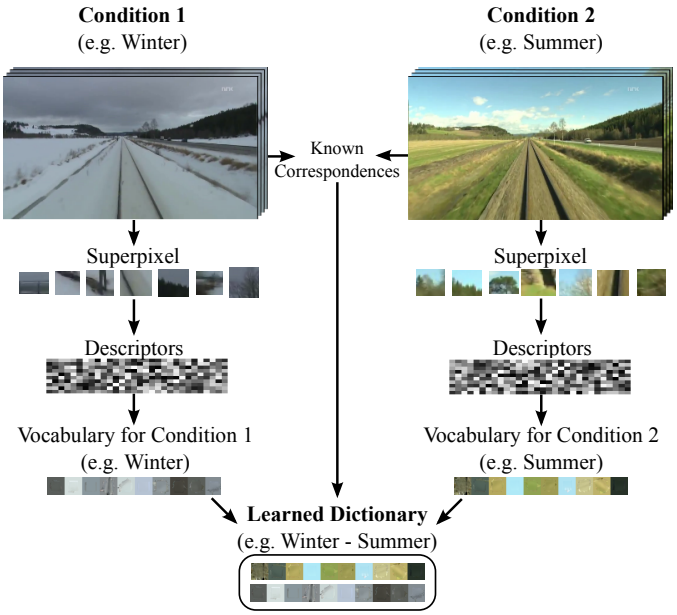

Fig. 2. Learning a dictionary between images under different environmental conditions (e.g. winter and summer). The images are first segmented into superpixels and a descriptor is calculated for each superpixel. These descriptors are then clustered to obtain a vocabulary of visual words for each condition. 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.

A. Learning Vocabularies and a Dictionary

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, e.g. captured by stationary webcams. Which approach to visual vocabulary learning is the most promising for the proposed scene change prediction has to be evaluated in future work.

Fig. 2 illustrates the training phase. In our current proof of concept implementation, each image is segmented into SLIC superpixels [1]. For each superpixel a descriptor that contains a color histogram in LAB color space (each channel with 10 bins), an U-SURF descriptor (128 byte) to capture texture information and the y-coordinate to encode spatial information 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.

Given the learned vocabularies, we can proceed to learn

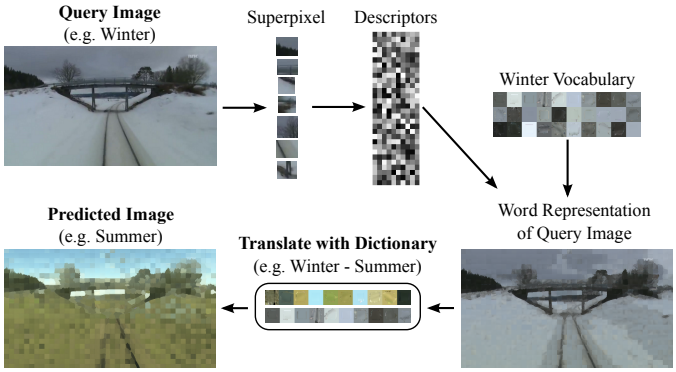

Fig. 3. Predicting the appearance of a query image under different environmental conditions: How would the current winter scene appear in summer? The query image is first segmented into superpixels and a descriptor is calculated for each of these segments. With this descriptor each superpixel can be 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. The result of the process is a synthesized image that predicts the appearance of the winter query image in summer time.

a dictionary that can translate between visual words from two environmental "languages" or conditions as illustrated in the lower part of Fig. 2. This dictionary captures the transitions of the visual words when the environmental conditions change. The dictionary can either capture the complete discrete probability distribution of these transitions or only store the transition that occurs most often.

B. 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.

If 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.

III. Experiments and Results

After the previous section explained how scene change prediction across seasons can be performed, we are going to describe the conducted experiments and their results.

A. The Nordland Dataset

To test our proposed approach of scene change prediction, we required a dataset where a camera traverses the same places under very different environmental conditions but under a similar viewing perspective: The TV documentary "Norlandsbanen – Minutt for Minutt" by the Norwegian Broadcasting Corporation NRK provides video footage of a 728 km long train ride that 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 such that an arbitrary frame from one video corresponds to the same frame of any of the other three videos etc. Therefore, frame-accurate ground truth information, e.g. corresponding scenes, are available. Furthermore, since the cameras were mounted exactly in the same spot in the driver's cabin, the four videos are almost perfectly aligned and thus allow easy learning of visual word transitions between the four seasons. The videos are available online at http://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/ under a Creative Commons licence (CC-BY).

For our experiments described in the following we extracted 30 minutes from the spring and the winter videos, starting approximately at 2 hours into the drive. From the four available videos, the spring video best resembled typical summer weather conditions. To form the training dataset, we extracted approximately 900 frames from the first 8 minutes of this 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 to evaluate the performance of the proposed scene change prediction.

B. Extending and Improving BRIEF-Gist and SeqSLAM

SeqSLAM [3] and BRIEF-Gist [4] are two established approaches to appearance-based place recognition. BRIEF-Gist is a holistic descriptor that encodes the visual appearance of a whole image in a short bit string. It supports place recognition by applying the Hamming distance between two descriptors in order to find the single global best matching query image. In contrast, SeqSLAM performs place recognition by matching whole *sequences* of images and has been shown to perform well despite severe appearance changes [3, 5]. We use OpenSeqSLAM [5] to perform the experiments.

Combining both approaches with our scene change prediction is particularly easy, since the change prediction algorithm can be executed as a preprocessing step before SeqSLAM or BRIEF-Gist start with their own processing. Since 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 BRIEF-Gist and SeqSLAM.

Fig. 4 compares precision-recall curves achieved by both algorithms with and without our proposed scene change prediction. The apparent result is that both BRIEF-Gist and SeqSLAM can immediately benefit from the change prediction. For SeqSLAM we plot the results for several values of the `ds` parameter that controls the minimal required length of the matched image sequences in seconds. We can see that SeqSLAM's performance increases with larger `ds`, as expected.

We can conclude that although SeqSLAM alone reaches good matching results, it can be significantly improved by first

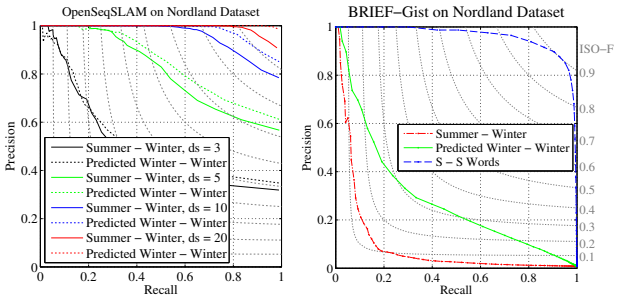

Fig. 4. Precision recall plots obtained by place recognition across seasons with SeqSLAM [3] (left) and BRIEF-Gist [4] (right). The plots compare the performance of the stand-alone algorithms with the boosted performance when the appearance of the winter images is predicted before place recognition is attempted. It is apparent that our proposed approach can significantly improve the performance of both algorithms. For comparison, the blue curve in the right plot shows the performance of BRIEF-Gist when matching summer images directly with summer word images, i.e. performing place recognition under constant environmental conditions.

predicting the appearance of the query scene under the viewing conditions of the stored database scenes. Also BRIEF-Gist can benefit from the proposed appearance change prediction, although its performance is in general much worse than SeqSLAM's.

For comparison we also evaluated the performance of FAB-MAP (using openFAB-MAP) on the dataset. As expected, directly matching winter against summer images was not successful. The best measured recall was 0.025 at 0.08 precision, presumably because FAB-MAP fails to detect common features in the images from both seasons.

IV. Discussion and Conclusions

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

We can synthesize an actual image during this 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 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 modelled and predicted. 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.

In its current form, our algorithm requires perfectly aligned images in the training phase. This requirement is hard to fulfill and limits the available training datasets. We will explore ways to ease this requirement in future work, e.g. by anchoring the training images on stable features. Another 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 that learns and maintains a single vocabulary that captures all conditions 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 and we therefore prepare to evaluate the proposed approach in a more general setting using data from vehicles in urban environments and training data that has been collected from stationary webcams over the course of several months.

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. volume 34, 2012.

[2] Winston Churchill and Paul M. Newman. Practice makes perfect? Managing and leveraging visual experiences for lifelong navigation. In *Proc. of Intl. Conf. on Robotics and Automation (ICRA)*, 2012.

[3] Michael Milford and Gordon F. Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *Proc. of Intl. Conf. on Robotics and Automation (ICRA)*, 2012.

[4] Niko Sünderhauf and Peter Protzel. BRIEF-Gist – Closing the Loop by Simple Means. In *Proc. of IEEE Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2011.

[5] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across All Four Seasons. In *Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)*, 2013.