Published at the Workshop on Long-Term Autonomy held in conjunction with the International Conference on Robotics and Automation (ICRA), 2013.

Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across All Four Seasons

Niko Sünderhauf, Peer Neubert, and Peter Protzel

***Abstract*— When operating over extended periods of time, an autonomous system will inevitably be faced with severe changes in the appearance of its environment. Coping with such changes is more and more in the focus of current robotics research. In this paper, we foster the development of robust place recognition algorithms in changing environments by describing a new dataset that was recorded during a 728 km long journey in spring, summer, fall, and winter. Approximately 40 hours of full-HD video cover extreme seasonal changes over almost 3000 km in both natural and man-made environments. Furthermore, accurate ground truth information are provided. To our knowledge, this is by far the largest SLAM dataset available at the moment. In addition, we introduce an open source Matlab implementation of the recently published SeqSLAM algorithm and make it available to the community. We benchmark SeqSLAM using the novel dataset and analyse the influence of important parameters and algorithmic steps.**

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.

While established state of the art approaches to place recognition like FAB-MAP [1] match *single* images, recent work of Milford and Wyeth [2] proposes to match *sequences* of images. Their approach, coined *SeqSLAM*, achieved quite remarkable results when recognizing places even if the environment underwent severe appearance changes, like transitioning from a sunny day to a rainy night.

Datasets that provide footage of changing environments in conjunction with high quality ground truth are currently rare, but absolutely crucial for the development and comparison of future place recognition algorithms. In the following, we therefore introduce a novel dataset and use it to analyse our open source implementation of SeqSLAM.

II. The Nordland Dataset

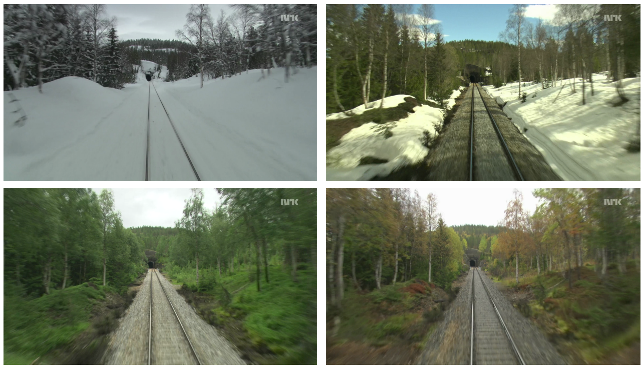


Fig. 1. The Nordland dataset consists of the video footage recorded on a 728 km long train ride in northern Norway. The journey was recorded four times, once in every season. Frame-accurate ground truth information makes this a perfect dataset to test place recognition algorithms under severe environmental 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/

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. 1, there is an immense variation of the appearance of the landscape, reaching from a complete snow-cover in winter over fresh and green vegetation in spring and summer to colored foliage in autumn. In addition to the seasonal change, different local weather conditions like sunshine, overcast skies, rain and snowfall are experienced on the long trip. Most of the journey leads through natural scenery, but the train also passes through urban areas along the way and occasionally stops at train stations or signals.

The videos 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 available online[1] under a Creative Commons licence (CC BY). The full-HD recordings have been time-synchronized 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

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

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

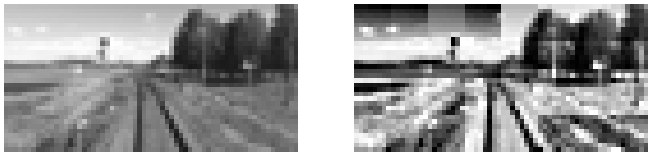


Fig. 2. SeqSLAM downsamples the original camera images to 64×32 pixels (left) and patch-normalizes them using patches of 8×8 pixels (right). The patch-normalized images are used to determine the image difference score which is calculated as the sum of absoulute differences of the pixel values.

positions and interpolating the GPS measurements to 25 Hz to match the video frame rate.

III. SeqSLAM and OpenSeqSLAM

SeqSLAM has been described in depth by Milford and Wyeth in their original publication [2]. For the sake of completeness we summarize the important ideas and algorithmic steps but refer the readers to [2] for the details.

In a preprocessing step, incoming images are drastically downsampled to e.g. 64×32 pixels. These thumbnail images are further divided into patches of 8×8 pixels, which are then normalized, so that the pixel values cover the complete range of possible values between 0 and 255. Fig. 2 illustrates this step of the algorithm. Given these patch-normalized thumbnails, an image difference matrix is calculated between all images using the sum of absolute differences.

The two important innovative steps of SeqSLAM are performed in the following: First, the distance matrix is locally contrast enhanced, which Milford and Wyeth describe as a step towards forcing the matcher to find best matches in every *local* neighborhood of the trajectory instead of only one *global* best match. Finally, when looking for a match to a query image, SeqSLAM performs a search to find the best matching *sequence* of adjacent frames. SeqSLAM literally sweeps through the contrast-enhanced difference matrix to achieve this.

In order to evaluate and possibly enhance SeqSLAM in the future, we wrote our own implementation in Matlab. The code uses the Parallel Computing Toolbox to parallelize large parts of the algorithm. We make this Matlab code available to the community on `www.openslam.org`.

IV. Experiments, Evaluation, and Results

For the experiments described in the following, we extracted still frames from the original videos using `avconv` on the Linux command line with the options `-r 1 -vsync vfr -s 64x32 -vf lutyuv="u=128:v=128"`. This extracts frames at 1 fps, downsamples them to 64×32 pixels and converts them into graylevel images on the fly. We provide the resulting images for download to the community. Details can be found in the `datasets/` directory in our release of OpenSeqSLAM.

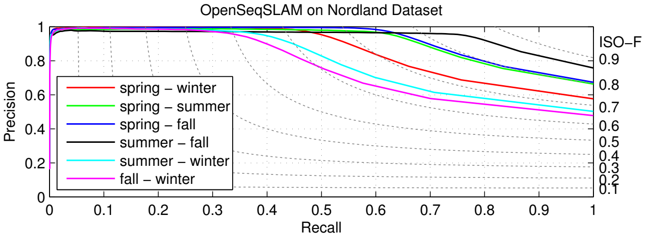


Fig. 3. General place recognition performance for the different season pairings, using sequence lengths of 10 s. Due to the individual characteristics of the seasonal changes, performance differs among the six possible pairings.

A. General Place Recognition Performance

In a first set of experiments, we evaluated how well OpenSeqSLAM is able to perform place recognition on the Nordland dataset in general. For this first test, we used a sequence length of 10 seconds (corresponding to 10 frames) and accepted a proposed match if it was within 1 frame of the ground truth. This is the default parameter setting for all experiments to follow, unless otherwise noted. Fig. 3 shows precision-recall plots for all possible six pairings among the seasons. The plots were created by varying SeqSLAM's parameter μ which controls the uniqueness of a sequence required for a match.

Despite the large variation in appearance and the seasonal changes, the results are remarkably good and reach high precision at reasonable recall rates. Since the recently developed robust SLAM back ends can cope with false positive loop closures, a precision of 80% may still be acceptable. Overall, the performance on the Nordland dataset is similar to the performance reported in [2] on two very different datasets.

B. Influence of the Sequence Length Parameter

SeqSLAM has a large number of free parameters that control the behaviour of the algorithm and have potential influence on the quality of the results. The temporal length of the image sequences that are considered for matching is controlled by the parameter d_s, which is presumably the most influential parameter. In general, one can expect SeqSLAM to perform better with longer sequence lengths in terms of precision and recall, since longer sequences are more distinct and thus less likely to result in false positive matchings. Fig. 4 illustrates this effect for a variety of d_s parameters. We see that sequence lengths of 20 or 30 seconds already give a very good performance on the Nordland dataset, while longer sequences do not show much improvement. Sequence lengths below 10 seconds lead to a drastic decrease in performance and should be avoided. Notice that for other datasets, longer sequence lengths can result in a lower recall, since the system will not recognize loop closures that overlap in temporal intervals shorter than d_s. This may be a problem especially in urban environments with frequent turns.

C. Contrast Enhancement and Patch Normalization

Milford and Wyeth [2] introduce the local contrast enhancement in the image difference matrix as a key innovation of their algorithm. They also propose to perform a patch normalization during the preprocessing, as illustrated in Fig. 2. We compare precision-recall curves with and without both individual steps in Fig. 5 to analyze their importance. The

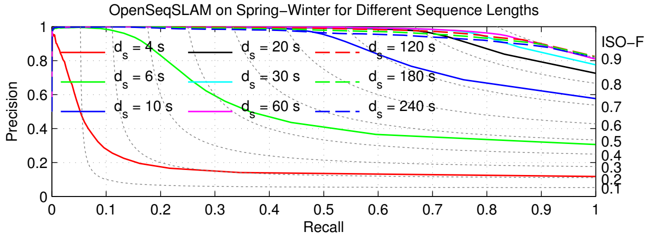


Fig. 4. Place recognition performance depending on the sequence length parameter d_s. A match was considered correct if it was within 1 second of the ground truth.

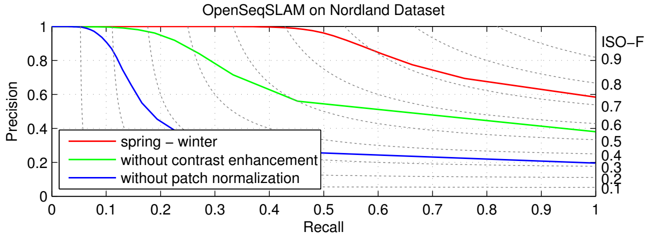


Fig. 5. Illustration of the influence of the contrast enhancement and patch normalization steps on the matching results. Experiments were performed on the spring - winter dataset. The red curve shows the default results and corresponds to the red curve in Fig. 3. For the blue and green curve, contrast enhancement and respectively patch normalization have been turned off.

influence of both processing steps is clearly visible, since the matching performance significantly drops when either one is disabled. The patch normalization however appears to have a much larger positive influence on the performance.

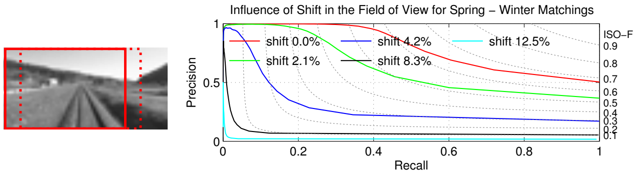
D. Field of View Dependence

A special feature of the Nordland dataset is that the viewpoint and the field of view in all four videos (i.e. in all four passes through the environment) is exactly the same. Since the camera was mounted in the same spot in the cockpit of the train and the train has only one degree of freedom along its track, corresponding images from two seasons overlap almost perfectly. Except for applications involving trains, this is not very realistic. We can usually expect a larger variation in viewpoints and fields of view between different traverses through an environment.

In order to determine how much the performance of SeqSLAM depends on the repeatability of the view point, we cropped the downsampled images of the spring and winter dataset further to 48×32 pixel. For the winter images we then shifted these downsampled crops by up to 6 pixels to the right. This way, we simulated a change in the field of view of up to 12.5% of the image resolution. Fig. 6(a) illustrates this experiment and Fig. 6(b) shows that the matching performance dramatically drops even for this comparably mild shift.

V. Conclusions

The last part of our experiments revealed a weakness of SeqSLAM: The surprisingly good performance depends mainly on the highly stable (in fact identical) viewpoint in the different traversals of the environment. To be more precise, it is a weakness of the algorithmic steps that calculate the difference between two images, i.e. patch normalization combined with sum of absolute difference. The core of SeqSLAM, namely the search for matching *sequences* is in principle unaffected by this effect. Thus it is absolutely worthwhile to combine image difference metrics that are invariant to reasonable viewpoint changes with the core of SeqSLAM in future work.



(a)

(b)

Fig. 6. The place recognition performance is extremely dependent on the field of view. (a) For this experiment, the 64×32 downsampled images were cropped to 48×32 (sold red frame). The crops for the winter images were then shifted by 0, 1, 2, 4, and 6 pixels to the right (dotted frame corresponds to a 6 pixel shift), giving rise to a slightly different field of view. This simulates the robot following a different trajectory through the environment. As plot (b) demonstrates, SeqSLAM's performance drops dramatically even with these relatively small shift values.

In summary, we feel that the presented Nordland dataset is a highly suitable test case for place recognition algorithms. However due to its special characteristics we strongly recommend that a test against view point invariancy is included in every paper that benchmarks an algorithm against this dataset. Otherwise the reported performance will most certainly be overly optimistic. OpenSeqSLAM provides a `crop` parameter that can be used to conveniently implement this test.

Acknowledgements: We sincerely thank the crew of NRKbeta.no for producing the Nordlandsbanen footage and making it available to the general public.

Discussion: 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, and SeqSLAM is among them. 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 an interesting topic for discussion.

References

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

[2] 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.