A neurologically inspired sequence processing model for mobile robot navigation

Stefan Schubert, Peer Neubert and Peter Protzel
Chemnitz University of Technology, Chemnitz, Germany
Email: {firstname.lastname}@etit.tu-chemnitz.de

I. Introduction

The current main goal of mobile robotics is to make robots more intelligent, and machine learning probably plays an import role to achieve this. Since a mobile robot moves in the environment, the difficulties of mobile robotics differ from those of stationary robotics: a mobile robot cannot return to a defined initial state, the proprioception does not return a global pose and the global pose is generally never exactly known, esp. visual perception is difficult to control in terms of lighting conditions or appearance changes, the closed world assumption does not apply as new and unknown relevant objects and situations could appear, and computational resources are somewhat limited due to weight, space and battery life. These differences are a subset of reasons why it is difficult to apply state-of-the-art deep reinforcement learning methods which have been successfully used in stationary robotics like sim-to-real transfer [22, 15] to visual navigation tasks in mobile robotics. Nevertheless, there were recent impressive results like learned visual navigation to an arbitrary target place presented by Bruce et al. [5]; however, in their work the agent was trained offline with reinforcement learning on images of a previously recorded environment – accordingly, online exploration of the environment is not possible with their method. These drawbacks of widely used algorithms motivate us to look for alternative machine learning approaches which could potentially be used in mobile robotics, and to figure out their limitations and potentials.

We focus on navigation as a fundamental ability of a mobile robot, and address the subtask *visual place recognition* in our current research. Visual place recognition is the problem of camera based localization of a robot given a database of images of known places, potentially under severe appearance changes (e.g., different weather or illumination) [16]. For instance, it is used for loop closure detection in order to build globally consistent maps in a SLAM system (Simultaneous Localization And Mapping) or to recover the robot's pose in case of tracking failure [20]. As places at different locations in the world can look similar, systems that exploit sequences of images (e.g., SeqSLAM [18]) potentially perform better than a pairwise image comparison [23].

To address the task of sequence-based visual place recognition in our current research, we take inspiration from the Hierarchical Temporal Memory (HTM) by Jeff Hawkins [10], a biologically plausible model of sequence processing in the human neocortex. HTM as a machine learning approach observes an incoming stream of sensor data and tries to learn to predict possible next inputs in an intermediate layer [11]; these predictions are based on previous (correct) predictions which enables the HTM system to encode long-term context even in case of temporarily similar input data but different context. In HTM Theory, input data and inner states are represented as binary SDRs (sparse distributed representations) [1], and learning is done in a Hebbian fashion. More details on HTM and current developments can be found in [9].

We took several ideas from HTM to develop our sequence-based visual place recognition algorithm *MCN* (MiniColumn Network) (see Sec. III). For performance evaluation, we conducted experiments on synthetic data, real world datasets, and online on a mobile robot in two challenging indoor environments (see Sec. IV). Building upon our current MCN-algorithm, we pursue ideas inspired by neurological insights to develop our system further for performance improvements (Sec. V), and to apply it to a broader area of visual navigation beyond place recognition.

II. Related Work

Place Recognition is a well studied problem. Lowry et al. [16] provide a recent survey. For place recognition in changing environments, descriptors from intermediate convolutional layers (e.g., conv3) from off-the-shelve CNNs like Alexnet [14] showed good results [23]; more recently, CNN descriptors were particularly designed and trained for place recognition, e.g. NetVLAD [2]. Based on such descriptors, a variety of approaches exists to compare and match images. Beyond simple pairwise comparison and using statistics of feature appearances (e.g., FAB-MAP [6]), the benefit of exploiting sequence information is well accepted: SeqSLAM [18] searches for linear segments of high similarity in the pairwise similarity matrix. Hansen and Browning [8] model sequence based place recognition as Hidden Markov Model. Arroyo et al. [3] use concatenated binary features to represent sequences. Vysotska et al. present a series of approaches to efficient place recognition using a graph theoretical approach [24, 25, 26]. RatSLAM [19] is an approach to SLAM that is biologically inspired by entorhinal grid cells [7] in the rats brain.

III. Approach

An early version of MCN is presented in [21] which we extended in our recent work for its application to real world data. A simple MCN is depicted in Fig. 1, right: It basically consists of a *spatial pooler* and a *temporal memory* which are inspired by HTM; the spatial pooler represents a feed-forward connection in the network from the input sensor data whereas the temporal pooler contains lateral connections to make predictions and to maintain an inner state representing

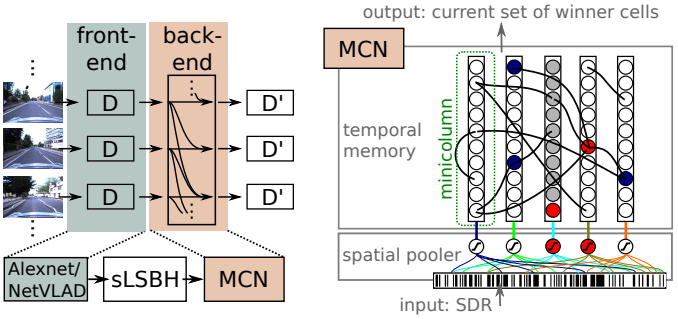

Fig. 1. *(left)* Overall approach. *(right)* Illustration of a minicolumn network (MCN). Red cells are current winner cells, grey cells are active (but not winner), blue cells are winner cells from the previous timestep.

the context of the input sequence. Like in HTM, the temporal memory is organized in minicolumns each containing multiple cells with directed connections between several cells of different minicolumns. Every minicolumn is sparsely connected over the spatial pooler with approx. 2% different elements of the current binary input SDR: If the number of connected 1s to a minicolumn exceeds a threshold t, this minicolumn is activated in case this activation belongs to the k_{max} most active minicolumns.

Every cell can have four different states: inactive, predicted, active, and winner. An inactive cell becomes predicted if at least one of its connected predecessor cells becomes active. A predicted cell becomes active if the corresponding minicolumn is activated by the spatial pooler in the next timestep; in this case all active cells become winner cells within this minicolumn. In conclusion, predicted cells in a minicolumn try to predict a potential activation of their minicolumn in the next timestep, so they try to predict a potential next input SDR to the MCN; if the minicolumn does not become active, the corresponding cells simply get inactive. However, if an activated minicolumn has no predicted cells, the minicolumn is bursted: all cells become active and one winner cell is chosen.

Learning in MCN is different to HTM and happens as one-shot learning as follows: Each winner cell is always additionally connected to all winner cells of the previous timestep. If the number of active minicolumns k_{act} is smaller than k_{min}, $k_{min} - k_{act}$ minicolumns are newly created with random connections to 1s in the current SDR (to get active in case of the same or similar input).

The data processing pipeline for place recognition is shown in Fig. 1, left: As front-end, an image descriptor of the current image is created with Alexnet or NetVLAD followed by a *sparse locality sensitive binary hashing* (sLSBH) that creates a high dimensional vector with 25% 1s. MCN serves as back-end to process the input stream sequentially.

Finally, all winner cells of a timestep represent the new descriptor of the current place, and are stored and compared to the sets of winner cells of every previous timestep to measure the similarity between the current and all previous places.

IV. Experiments

We performed experiments with MCN on synthetic data, real-world datasets, and online on a mobile robot: As MCN has a couple of parameters, we evaluated the place recognition

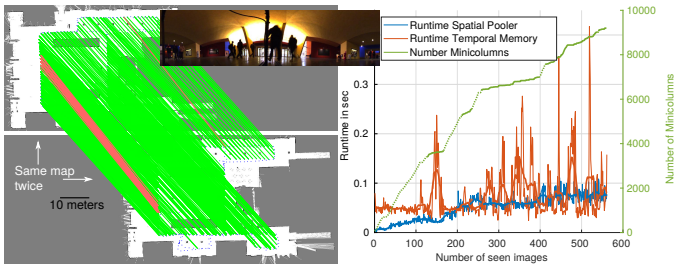

Fig. 2. Online experiment. Thick curves are 21-frame sliding averages of thin curves. Dotted green line indicates exploration, solid line revisits.

performance (*maxF-score*) as a function of each parameter. We found that higher values are always better regarding maxF-score (potentially with the cost of higher computation effort), except for the activation threshold t which has to be set more carefully.

Furthermore, we tested MCN with the two different front-ends Alexnet and NetVLAD on five different real-world datasets (with 22 different sequence-combinations) like Oxford RobotCar [17] or CMU [4] dataset each with different challenging properties like multiple loop closures within one sequence, zero-speed, and appearance changes due to weather, season, daytime or dynamic objects. We compared our results to six different sequence processing back-ends [18, 3, 25, 24, 26, 8]. Considering average precision, MCN performs best on more sequences than any other algorithm; furthermore, it is the only algorithm which never performs much worse than the pairwise image comparison baseline – accordingly, MCN always maintains or improves its front-end's performance. NetVLAD as front-end performs better than Alexnet.

Finally, as a proof-of-concept we ran MCN online on a mobile robot equipped with a camera to detect loop closures. The robot drove multiple loops in two different indoor environments both with a modern, clean architecture. Fig. 2 shows the result of a 730m long ride through the foyer of a lecture hall building with passing students. The system achieved 858 true and 33 false loop closures. Note that the robot performed visual place recognition but not visual localization; metric information is used only for visualization of ground truth. The plot on the right shows that the number of minicolumns grows much slower in case of revisits. We also achieved good results for the second environment.

V. Discussion & Future Work

We could achieve first good results with MCN on many different experiments. However, the system requires more theoretical analysis as well as further investigation of the prediction quality and length within the temporal memory. Furthermore, we want to involve more concepts from the original HTM like permanences and segments. To extend MCN's capabilities and to perform more sophisticated navigation tasks, we are going to extend the system with other cell types like grid cells or head-direction cells [7] as proposed in [12, 13], or to figure out how multimodal data like camera and LiDAR can be used. A distant goal might be the output of actions, but this requires further extensive research.

References

[1] Subutai Ahmad and Jeff Hawkins. Properties of sparse distributed representations and their application to hierarchical temporal memory. *CoRR*, abs/1503.07469, 2015.

[2] R. Arandjelovi, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *Trans. on Pattern Analysis and Machine Intelligence*, 40(6), 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2711011.

[3] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera. Towards life-long visual localization using an efficient matching of binary sequences from images. In *Proc. of Int. Conf. on Robotics and Automation*, 2015. doi: 10.1109/ICRA.2015.7140088.

[4] H. Badino, D. Huber, and T. Kanade. Visual topometric localization. In *Proc. of Intelligent Vehicles Symp.*, 2011.

[5] Jake Bruce, Niko Sünderhauf, Piotr W. Mirowski, Raia Hadsell, and Michael Milford. Learning deployable navigation policies at kilometer scale from a single traversal. In *Conference on Robot Learning (CoRL)*, 2018.

[6] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The Int. J. of Robotics Research*, 27(6), 2008.

[7] Roddy M. Grieves and Kate J. Jeffery. The representation of space in the brain. *Behavioural Processes*, 135:113 – 131, 2017.

[8] P. Hansen and B. Browning. Visual place recognition using hmm sequence matching. In *Proc. of Int. Conf. on Intelligent Robots and Systems*, 2014. doi: 10.1109/IROS.2014.6943207.

[9] J. Hawkins, S. Ahmad, S. Purdy, and A. Lavin. Biological and machine intelligence (bami). Initial online release 0.4, 2016. URL https://numenta.com/resources/biological-and-machine-intelligence/.

[10] Jeff Hawkins. *On Intelligence (with Sandra Blakeslee)*. Times Books, 2004.

[11] Jeff Hawkins and Subutai Ahmad. Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Frontiers in Neural Circuits*, 10:23, 2016. ISSN 1662-5110. doi: 10.3389/fncir.2016.00023.

[12] Jeff Hawkins, Subutai Ahmad, and Yuwei Cui. A theory of how columns in the neocortex enable learning the structure of the world. *Frontiers in Neural Circuits*, 11: 81, 2017. ISSN 1662-5110. doi: 10.3389/fncir.2017.00081.

[13] Jeff Hawkins, Marcus Lewis, Scot Purdy, Mirko Klukas, and Subutai Ahmad. A framework for intelligence and cortical function based on grid cells in the neocortex. *Frontiers in Neural Circuits*, 12, 2019. doi: 10.3389/fncir.2018.00121.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 2012.

[15] Robert Lee, Serena Mou, Vibhavari Dasagi, Jake Bruce, Jürgen Leitner, and Niko Sünderhauf. Zero-shot sim-to-real transfer with modular priors. *CoRR*, abs/1809.07480, 2018.

[16] Stephanie Lowry, Niko Sunderhauf, Paul Newman, John J. Leonard, David Cox, Peter Corke, and Michael J. Milford. Visual place recognition: A survey. *Trans. Rob.*, 32(1), 2016. ISSN 1552-3098. doi: 10.1109/TRO.2015.2496823.

[17] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The Int. J. of Robotics Research*, 36(1):3–15, 2017. doi: 10.1177/0278364916679498.

[18] Michael Milford and Gordon Fraser Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *Proc. of Int. Conf. on Robotics and Automation*, 2012. ISBN 978-1-4673-1403-9.

[19] Michael Milford, Gordon Wyeth, and David Prasser. Ratslam: a hippocampal model for simultaneous localization and mapping. In *Proc. of Int. Conf. on Robotics and Automation*, 2004.

[20] Montiel J. M. M. Mur-Artal, Raúl and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. doi: 10.1109/TRO.2015.2463671.

[21] Peer Neubert, Subutai Ahmad, and Peter Protzel. A sequence-based neuronal model for mobile robot localization. In *Proc. of KI: Advances in Artificial Intelligence*, 2018.

[22] Andrei A. Rusu, Matej Vecerik, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. In *CoRL*, 2017.

[23] Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. On the performance of convnet features for place recognition. *Proc. of Int. Conf. on Intelligent Robots and Systems*, 2015.

[24] O. Vysotska and C. Stachniss. Lazy data association for image sequences matching under substantial appearance changes. In *IEEE Robotics and Automation Letters*, volume 1, 2016. doi: 10.1109/LRA.2015.2512936.

[25] O. Vysotska, T. Naseer, L. Spinello, W. Burgard, and C. Stachniss. Efficient and effective matching of image sequences under substantial appearance changes exploiting gps priors. In *Proc. of Int. Conf. on Robotics and Automation*, 2015. doi: 10.1109/ICRA.2015.7139576.

[26] Olga Vysotska and Cyrill Stachniss. Relocalization under substantial appearance changes using hashing. In *Proc. of 9th Workshop on Planning, Perception, and Navigation for Intelligent Vehicles at the Int. Conf. on Intelligent Robots and Systems*, 2017.