DOI: 10.48011/asba.v2i1.1036

Long-Term Visual Robot Localization using Computational Models of the Neocortex

Carlos Alexandre P. Pizzino * **Patricia A. Vargas** **
Ramon R. Costa ***

* *Department of Electrical Engineering, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, (e-mail: pizzino@ufrj.br)*
** *Edinburgh Centre for Robotics, Heriot-Watt University, Edinburgh, Scotland, United Kingdom, (e-mail: p.a.vargas@hw.ac.uk)*
*** *Department of Electrical Engineering, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, (e-mail: ramon@coep.ufrj.br)*

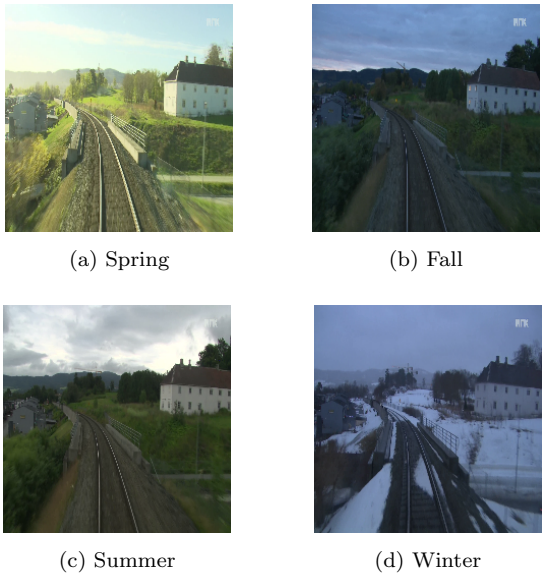
Abstract:
Visual place recognition is an essential capability for autonomous mobile robots which use cameras as their primary sensors. Although there has been a considerable amount of research in the topic, the high degree of image variability poses extra research challenges. Following advances in neuroscience, new biologically inspired models have been developed. Inspired by the human neocortex, hierarchical temporal memory model has potential to identify temporal sequences of spatial patterns using sparse distributed representations, which are known to have high representational capacity and high tolerance to noise. These features are interesting for place recognition applications. Some authors have proposed simplifications from the original framework, such as starting from an empty set of minicolumns and increasing the number of minicolumns on demand instead of the usage of a fixed number of minicolumns whose connections adapt over time. In this paper, we investigate the usage of framework originally proposed with the aim of extending the run-time during long-term operations. Results show that the proposed architecture can encode an internal representation of the world using a fixed number of cells in order to improve system scalability.

Keywords: Visual Place Recognition, Biologically-Inspired Robots, Spatial Representations in the Brain, Hebbian Learning, Hierarchical Temporal Memory.

1. INTRODUCTION

Autonomous mobile robots can be used in many challenging applications such as environment reconstruction, search and rescue, underwater surveillance, autonomous driving and planetary exploration. In order to be able to succeed in such complex scenarios, Thrun et al. (2005) and Stachniss et al. (2016) state that one of the main capabilities of a truly autonomous mobile robot is that of being able to perform concurrently mapping and tracking of its own location, also know as Simultaneous Localization and Mapping (SLAM).

To perform SLAM, the ability to recognize a previously visited place is a fundamental task (Sünderhauf et al., 2015). Therefore, place recognition systems must build and keep a consistent map of the environment by comparing incoming data with representations that are already included in the map. Nevertheless, visual place recognition in *long-term* and *large-scale* environments is a rather challenging problem. In these scenarios the robot must take into account the context of image variations caused by different times of the day, weather, lighting and seasons conditions (Figure 1).



(a) Spring (b) Fall

(c) Summer (d) Winter

Figure 1. Examples of different seasons conditions that might affect place recognition (Sünderhauf et al., 2013). Source: Nordland dataset Skrede (2013).

The visual place recognition system can be divided into three important phases (Lowry et al., 2016). Firstly, it is necessary to able to describe places. The most common approaches are based on local image features using classic extractors and descriptors such as SIFT (Lowe, 2004), SURF (Bay et al., 2008), and GIST (Oliva and Torralba, 2001). However, these methods rely on purely hand-crafted features and are not appropriate for dealing with strong visual changes. Therefore, the most recent approaches use Convolutional Neural Networks (CNN), due to their higher potential to learn image patterns (Ranieri et al., 2020; Ferreira et al., 2017). Secondly, the system must maintain a stored representation of the robot's knowledge of the world, with or without associated position information (pure image retrieval). Finally, it is necessary to determine whether a place has been seen before by defining a similarity function (Carvalho et al., 2019). Most successful approaches explore sequential information, as there is a high correlation within consecutive inputs and across similar trajectories.

Hierarchical Temporal Memory (HTM) is a nice example of system that is capable of extracting spatial and temporal patterns from a time varying input data. HTM is an established brain inspired model of working principles of the human neocortex proposed by Hawkins et al. (2016). Following recent advances in neuroscience findings, new models have been proposed that are based on HTM with many applications in robotics. For instance, Neubert et al. (2018) explored neurally inspired approach for place recognition based on minicolumn network (MCN), which is a simplified version of HTM. MCN creates an internal representation that encodes sequential context through binary sparse distributed representations. This approach was extended in Neubert et al. (2019), where each image is processed by CNN-based descriptors. Their encoding is a sparsified binary adaptation of locality sensitive hashing (LSH) based on random projections. However, several simplifications were made. One of them was the creation of new minicolumns for unseen observation instead of using a fixed set of minicolumns. During exploration of new areas, the algorithm creates new columns, which can affect the run-time, so the maximum number of minicolumns should be limited. Another simplification includes the absence of a spatial pooler and segments, and the usage of one-shot learning instead of Hebbian-like learning.

In this work, we investigate the usage of framework originally proposed by Hawkins and Blakeslee (2004). In this model, learning involves incrementing or decrementing the permanence values of potential synapses on a dendrite segment. Preliminary results show that the proposed architecture can encode an internal representation of the world using a fixed number of minicolumns and cells for long-term SLAM. This work only stores appearance information about each place in the environment, with no associated position information

2. VISUAL PLACE RECOGNITION

In the past several years, visual place recognition in *long-term* and *large-scale* environment has been intensively investigated. Cummins and Newman (2008) introduced the FAB-MAP algorithm, a probabilistic framework for navigation and mapping which relies on appearance information only, and Cummins and Newman (2011) modified the structure of original model to support very large scale place recognition. Milford and Wyeth (2012) proposed the SeqSLAM, a sequence based place recognition algorithm, which adopts sequence matching rather than single image matching and achieves significant performance improvements to lighting changes. Arroyo et al. (2015) presented the ABLE-S, an approach based on binary codes and disparity information and Naseer et al. (2015) presented an appearance-based visual SLAM approach. Along these lines, Lowry et al. (2016) provides a recent survey based on image processing front-ends, of which a variety of approaches exists to compare and match images. Estimating a current pose can be used to support visual place recognition as well.

Sünderhauf et al. (2015) investigates the use of Convolutional Network (ConvNet) features in visual navigation and SLAM. They demonstrated that *large-scale* robust place recognition using ConvNet features is possible when applying a specialized binary hashing method. They evaluated their experiments based on single image matching performance, which can be a limiting factor for *large-scale* place recognition. The nearest neighbor search run-time is proportional to the number of stored previously visited places.

Another example of a visual place recognition system is the usage of Hierarchical Temporal Memory (HTM) that models a neural network by drawing inspirations from human neocortex Hawkins and Blakeslee (2004). According to the researchers, the neocortex is the part of the human brain that is involved in higher-order functions such as conscious thought, spatial reasoning, language, generation of motor commands, and sensory perception. Zhang et al. (2012) investigated a mapping strategy based on a HTM model in the form of a 2-level hierarchical structure and obtaining visual words from SURF descriptors, the mapping problem was treating as an object recognition problem. Mai et al. (2013) proposed a strategy to explore perception-action, based on HTM. They implemented perception-action system by perceiving images. In Pal et al. (2018), the authors model the deviation in learning for HTM when applied to a robotic path learning scenario.

Zhang et al. (2018) explored the sequential perceptual information combined with motion data simultaneously, which contributes to predicting one-step future actions. Basically, the system perceives images and incorporates with depth and motion command data to encode as a sequence of sparse distributed representation vectors. The sequential vectors are inputs to train the navigation hierarchical temporal memory. After the training, the the network stores the transitions of the perceived images, depth, and motion data so that future motion commands can be predicted.

Neubert et al. (2018) introduced a simplified version of HTM, called Simplified Higher Order Sequence Memory (SHOSM), which some simplifications were made, including the absence of a spatial pooler and segments, and the usage of one-shot learning instead of Hebbian-like learning. This approach creates an internal representation that encodes sequential context, using Sparse Distributed Repre-

sentations (SDRs), which are known to have high representational capacity and high robustness towards noise. They discussed interesting theoretical association of aspects of HTM theory and the problem of mobile robot localization. Neubert et al. (2019) extended the previous theoretical work and applied it to real world data, in particular, in combination with CNN-based image descriptors. The images were encoded with a CNN-based network like AlexNet (Krizhevsky et al., 2012) and NetVLAD (Arandjelović et al., 2018), and binarized with sparse Locality Sensitive Binary Hashing (sLSBH) based on random projections. The sparsified binary is the input to the Minicolumn Network (MCN), which is a Simplified Higher Order Memory, and encodes an internal representation based on sequential context.

Our proposed model differs from the previous models for it is composed of HTM Cortical Learning Algorithms and it will be further described in the next Section.

3. HIERARCHICAL TEMPORAL MEMORY MODEL

Hawkins and Blakeslee (2004) proposed a biologically inspired theoretical framework capable of human brain-like learning and prediction. This technology is known as Hierarchical Temporal Memory (HTM) that captures the structural and algorithmic properties of the neocortex. The HTM model is composed of numerous interconnected cells, which are organized in a column paradigm.

HTM has two distinct components namely Spatial Pooler (SP) and Temporal Memory (TM). The spatial pooler is responsible for creating the SDRs for the given spatio-temporal input data patterns. The spatial pooler algorithm follows unsupervised learning and uses vector quantization to perform its operations. The temporal memory of the HTM, on the other hand, is responsible for making predictions using the learned sequences for the given input. The working principle of temporal memory component is similar to Hebb's rule. Learning an input sequence involves formation of connections between cells within the active columns in the region.

The model works based on the principles of Sparse Distribution Representation (SDR). This representation is an encoding technique in which only a small percentage of the overall bits are active at a time. This encoding process could be loose information but this loss is often negligible since the loss will not have a substantial effect (Hawkins et al., 2016). The usage of the SDRs also reduces the memory needed to store the data and the power consumption in the hardware.

HTM models learn continuously, which is often referred to as on-line learning. With each change in the inputs the memory of the HTM system is updated. There are no batch learning data sets and no batch testing sets as is the norm for most machine learning algorithms. Another advantage of continuous learning is that the system will constantly adapt if the patterns in the world change. For a biological organism this is essential to survival.

Differently from Neubert et al. (2019), this paper implement a HTM algorithms for learning and prediction developed by Numenta (2018). The main advantage of our approach is scalability for the robot can run during *long-term* operations, as new minicolumns are not created during exploration of new areas.

4. EXPERIMENTS AND RESULTS

In the section, we compare and show the results of our approach compared to the algorithm proposed in Neubert et al. (2018) and Neubert et al. (2019). In the experiments, a standard laptop with an i7-6500U CPU @ 2.50GHz was used. They were implemented in Python 2.7.17 using Numenta (2018) Platform for Intelligent Computing (NuPIC) platform, for HTM implementation. CNN features from conv3 layer were computed by a pre-trained AlexNet using Pytorch.

To evaluate the results, we use Precision-Recall (PR) curve, which is the most popular evaluation approach of place recognition algorithm. The precision-recall curve shows the agreement between precision and recall for different threshold. While high precision relates to a low false positive rate, high recall relates to a low false negative rate. In the words, a high area under the curve represents both high recall and high precision (Area Under the Curve - AUC).

4.1 Single Grid-Like World

This section compares MCN to HTM approach under controlled conditions. We use the same simulation environment proposed by Neubert et al. (2018), which simulates a path in a grid-like world.

The robot's sensor provides a 2,048 dimensional (40 1-bits, i.e. roughly 2% of sparsity SDR for each grid cell), which is considered a place. The robot path is a sequence of neighbored places. We use the same trajectory, and vary the amount of observation noise and place-aliasing. More precisely, the noise parameter controls the ratio of 1-bits that are erroneously moved in the observed SDR (n), while the place-aliasing parameter (a) counts the number of pairs of places in the world which look exactly the same.

In our implementation, we do not use Spatial Pooler, as the robot's sensor creates SDRs and perform vector quantization process. The number of cell per column is 4, and are distributed within 2048 columns in Temporal Memory.

Using the ground-truth information, precision and recall are computed. The results of MCN, HTM and pairwise comparison for different amount of aliasing and observation noise can be seen in Figures 2 and 3.

The Figure 4 illustrates the similarity matrices for a place recognition experiment with 11 loops. It is important to note that each entry is the similarity of a current query image to a dataset image. On the right side, the similarities are obtained from overlap of the sparse vector of winner cells.

MCN and HTM models benefit from the usage of sequential information and perform better than pairwise comparison.

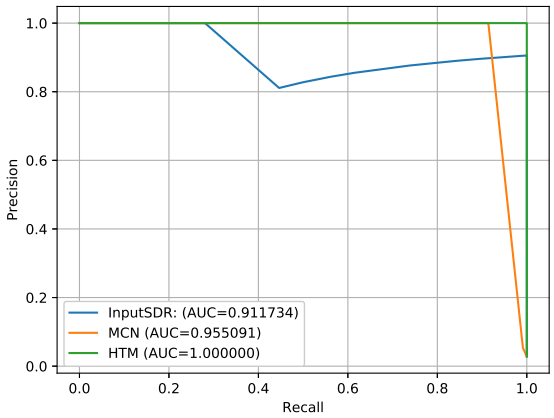


Figure 2. Precision-Recall curve for $a = 0$ and $n = 1\%$.

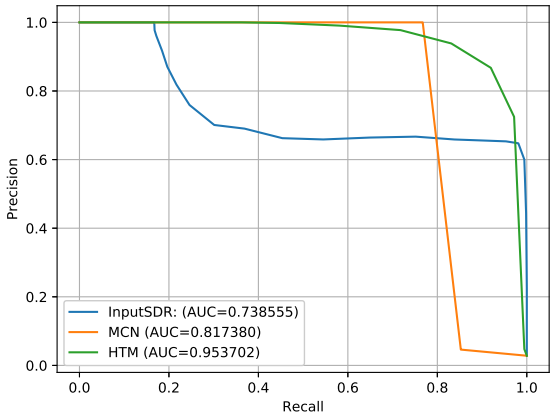


Figure 3. Precision-Recall curve for $a = 5$ and $n = 5\%$.

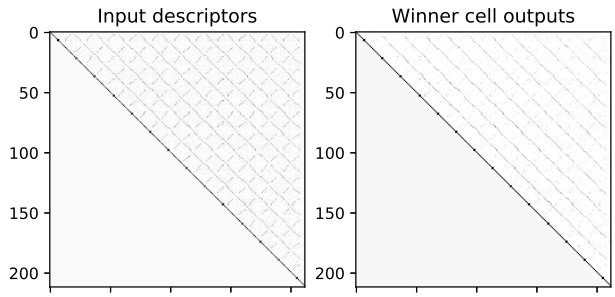


Figure 4. Distance matrices obtained by HTM algorithm ($a = 5$ and $n = 5\%$).

4.2 Nordland Dataset

The Nordland dataset is composed of images of all seasons from four journeys on a 728 km train route across Norway. In this work, we use the Nordland dataset partitioned by Olid et al. (2018).

The Figure 5 illustrates the overall approach and has similarities with the work proposed by Neubert et al. (2019). Each image is processed by a visual front-end that extracts a descriptor. Sünderhauf et al. (2015) studied the performance of features from different neural networks for the purpose of place recognition. In our case, we use only the output of the conv3 layer of AlexNet, which descriptor has $n = 64,896$ dimensions. However, both HTM and MCN use SDR inputs. For this, we use the algorithm proposed by Neubert et al. (2019), which is a sparsified binary adaptation of locality sensitive hashing (LSH) based on random projections (sLSBH - Sparsified Binary LSH).

The experimental results of AlexNet+sLSBH+MCN and AlexNet+sLSBH+HTM are shown as follows (Figure 6, 7 and 8). For each simulation, we compute precision-recall curve and report Area Under the Curve (AUC) using the trapezoidal rule. We use a pairwise comparison of individual descriptors using cosine distance for AlexNet and Hamming distance for sLSBH algorithm.

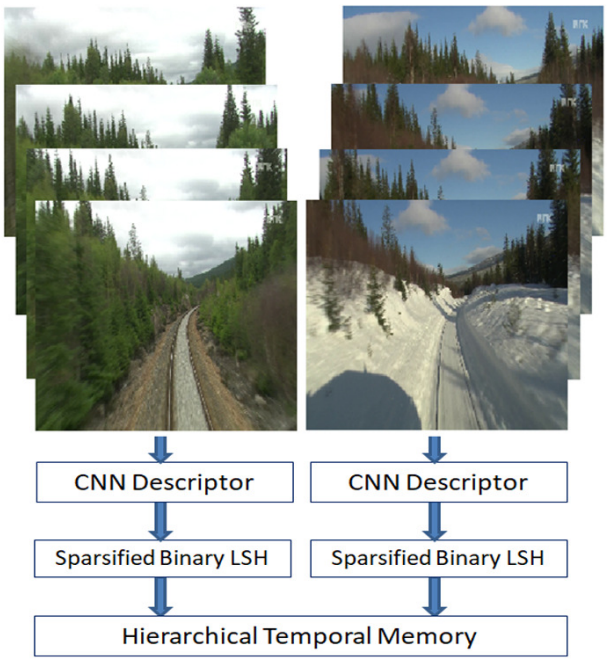


Figure 5. Images from the same place in different seasons. The images have been extracted from the videos of the Nordland dataset. General diagram life-long visual topological localization proposed.

Both HTM and MCN are trained on the database sequence (DB). The output winner cell descriptors for each image are store in a database and descriptors of new images are compared to the database using overlap metric. In our simulations, the number of seen images is 200.

In these simulations, the HTM has a slightly higher AUC (Area Under the Curve) than MCN. Moreover, HTM has a fixed number of minicolumns, whereas, in the MCN, new minicolumns are created during exploration of new areas. The curves on the right side of figures evaluate the growth of the MCN system.

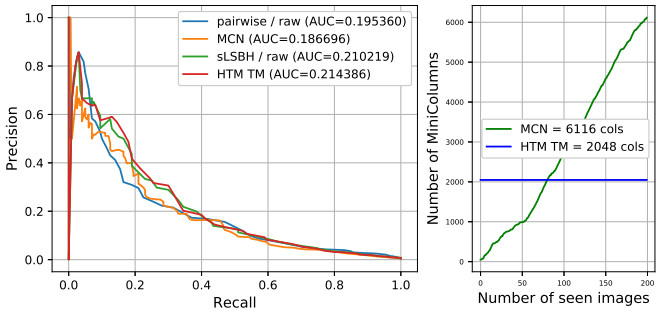


Figure 6. Comparison of Precision-Recall curves: summer (DB) and winter (query) seasons.

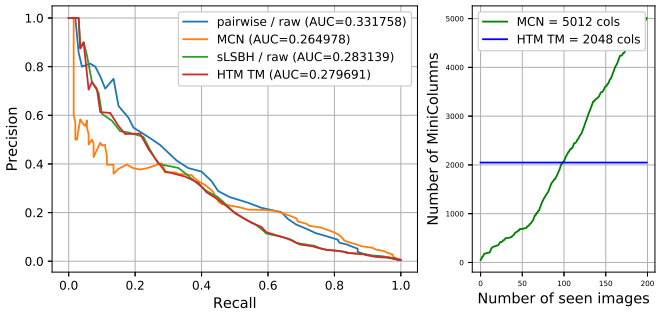


Figure 7. Comparison of Precision-Recall curves: spring (DB) and fall (query).

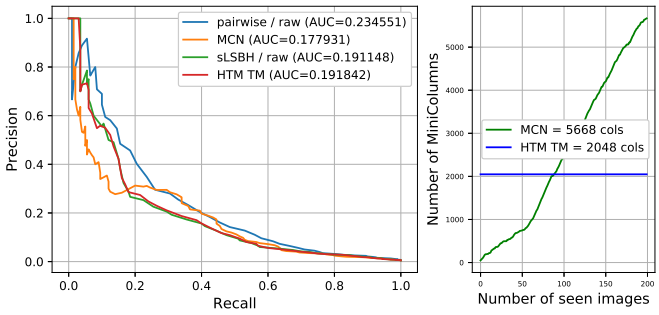


Figure 8. Comparison of Precision-Recall curves: fall (DB) and winter (query).

5. CONCLUSIONS AND FUTURE WORK

In this work we proposed and implemented a new visual place recognition system using a Hierarchical Temporal Memory (HTM) model. HTM is originally based on neuroscience research focused on the structure of the neocortex using minicolumn networks of neurons. The first HTM's implementation for mobile robot localization had several simplifications that included the absence of a spatial pooler and segments, and the usage of one-shot learning instead of Hebbian-like learning. Moreover, the creation of a new minicolumn for each new observed pattern jeopardised the run-time capability of the robot.

Our proposal has fixed number of minicolumns, whose connections adapt over time. Therefore, the main advantage of our system is *scalability*. The robot can run for *long-term* operations without an uncontrolled explosion on the number of new minicolumns, hence the run-time is completely scalable.

We evaluated two simulation scenarios. The results obtained were satisfactory, and confirm the usage of HTM for place recognition. As can be seen, the number of columns in the MCN approach increases over time. There are several parameters of HTM that must be analyzed in the future. Besides, the application on real data images should be investigated as this is a problem of suitable SDR encoders for typical robot sensors like cameras and laser scanners.

REFERENCES

Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2018). Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6), 1437–1451.

Arroyo, R., Alcantarilla, P.F., Bergasa, L.M., and Romera, E. (2015). Towards life-long visual localization using an efficient matching of binary sequences from images. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 6328–6335.

Bay, H., Ess, A., Tuytelaars, T., and Gool], L.V. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3), 346 – 359. doi:https://doi.org/10.1016/j.cviu.2007.09.014. URL `http://www.sciencedirect.com/science/article/pii/S1077314207001555`. Similarity Matching in Computer Vision and Multimedia.

Carvalho, E.C., Ferreira, B.V., Filho, G.P.R., Gomes, P.H., Freitas, G.M., Vargas, P.A., Ueyama, J., and Pessin, G. (2019). Towards a smart fault tolerant indoor localization system through recurrent neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–7.

Cummins, M. and Newman, P. (2008). Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6), 647–665. doi:10.1177/0278364908090961. URL `https://doi.org/10.1177/0278364908090961`.

Cummins, M. and Newman, P. (2011). Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research*, 30(9), 1100–1123. doi:10.1177/0278364910385483. URL `https://doi.org/10.1177/0278364910385483`.

Ferreira, B.V., Carvalho, E., Ferreira, M.R., Vargas, P.A., Ueyama, J., and Pessin, G. (2017). Exploiting the use of convolutional neural networks for localization in indoor environments. *Applied Artificial Intelligence*, 31(3), 279–287. doi:10.1080/08839514.2017.1316592.

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

Hawkins, J. and Blakeslee, S. (2004). *On Intelligence*. Times Books, USA.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25. doi:10.1145/3065386.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 91–110. doi:10.1023/B%3AVISI.0000029664.99615.94.

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J.J., Cox, D., Corke, P., and Milford, M.J. (2016). Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1), 1–19. doi:10.1109/TRO.2015.2496823.

Mai, X., Zhang, X., Jin, Y., Yang, Y., and Zhang, J. (2013). Simple perception-action strategy based on hierarchical temporal memory. In *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 1759–1764. doi:10.1109/ROBIO.2013.6739722.

Milford, M.J. and Wyeth, G.F. (2012). Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *2012 IEEE International Conference on Robotics and Automation*, 1643–1649.

Naseer, T., Ruhnke, M., Stachniss, C., Spinello, L., and Burgard, W. (2015). Robust visual slam across seasons. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2529–2535.

Neubert, P., Schubert, S., and Protzel, P. (2019). A neurologically inspired sequence processing model for mobile robot place recognition. *IEEE Robotics and Automation Letters*, 4(4), 3200–3207. doi:10.1109/LRA.2019.2927096.

Neubert, P., Ahmad, S., and Protzel, P. (2018). A sequence-based neuronal model for mobile robot localization. In F. Trollmann and A.Y. Turhan (eds.), *KI 2018: Advances in Artificial Intelligence*, 117–130. Springer International Publishing, Cham.

Numenta (2018). Nupic - numenta platform for intelligent computing. `https://github.com/numenta/nupic`. doi:10.5281/zenodo.1257382. Accessed: 2020-06-20.

Olid, D., Fácil, J.M., and Civera, J. (2018). Single-view place recognition under seasonal changes. In *PPNIV Workshop at IROS 2018*.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42, 145–175. doi:10.1023/A:1011139631724.

Pal, K., Bhattacharya, S., Dey, S., and Mukherjee, A. (2018). Modelling htm learning and prediction for robotic path-learning. In *2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob)*, 839–844. doi:10.1109/BIOROB.2018.8487228.

Ranieri, C.M., Vargas, P.A., and Romero, R.A.F. (2020). Uncovering human multimodal activity recognition with a deep learning approach. In *To appear in the Proc. of the 2020 International Joint Conference on Neural Networks (IJCNN)*.

Skrede, S. (2013). Nordlandsbanen: minute by minute, season by season. `https://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/`. Accessed: 2020-06-21.

Stachniss, C., Leonard, J.J., and Thrun, S. (2016). *Simultaneous Localization and Mapping*. Springer International Publishing, Cham. doi:10.1007/978-3-319-32552-1_46. URL `https://doi.org/10.1007/978-3-319-32552-1_46`.

Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., and Milford, M. (2015). On the performance of convnet features for place recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4297–4304.

Sünderhauf, N., Neubert, P., and Protzel, P. (2013). Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. *Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press.

Zhang, X., Zhang, J., Rad, A.B., Mai, X., and Jin, Y. (2012). A novel mapping strategy based on neocortex model: Pre-liminary results by hierarchical temporal memory. In *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 476–481. doi:10.1109/ROBIO.2012.6491012.

Zhang, X., Zhang, J., and Zhong, J. (2018). Toward navigation ability for autonomous mobile robots with learning from demonstration paradigm: A view of hierarchical temporal memory. *International Journal of Advanced Robotic Systems*, 15(3), 1729881418777939. doi:10.1177/1729881418777939. URL `https://doi.org/10.1177/1729881418777939`.