S2L-SLAM: Sensor Fusion Driven SLAM using Sonar, LiDAR and Deep Neural Networks

Niels Balemans*†, Peter Hellinckx*, Steven Latré*, Philippe Reiter* and Jan Steckel†‡
Email: {firstname}.{lastname}@uantwerpen.be

2021 IEEE Sensors | 978-1-7281-9501-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/SENSORS47087.2021.9639772

***Abstract*—The use of different modalities improves the perception of the environment in situations where the conventional sensors fail (camera and LiDAR). The inclusion of these modalities, such as sonar or radar, is however difficult as existing methods for the conventional sensors usually do not generalise well on these different environment representations. We experiment with a modality prediction method to keep using the existing methodologies and allow to separate the sensing system from the navigation stack of an autonomous agent. In previous work, we used a convolutional stacked autoencoder to predict LiDAR point cloud data using the data from our 3D in-air acoustic ultrasonic sensor (eRTIS). In this paper, we investigate the usability of the predicted data in off-the-shelf algorithms to safely navigate environments where visual modalities become unreliable and less accurate.**

***Index Terms*—Sensor systems, Ultrasonic measurements, Inverse problems, Navigation, Deep learning**

I. Introduction

The usage of autonomous vehicles has already shown its potentials in many different domains ranging from conventional consumer cars to industrial applications [1], [2], [3], [4]. It is evident that for the safe and reliable deployment of these systems, the perception of the vehicle's environment is immensely important. The conventional state-of-the-art sensing setups use camera and time-of-flight (ToF) sensors (i.e. LiDAR) to map their surroundings and detect obstacles and potential safety hazards [5], [6]. This means that many of the existing algorithms are tailored for camera and LiDAR and often are not easily applied to other sensing modalities. We argue, however, that the use of different sensing modalities, i.e. sonar and radar, will greatly benefit the environment perception in harsh sensing conditions. Being able to accurately and robustly perceive an environment, even in difficult conditions, will not only extend the application potential but also increase the safety of autonomous agents. The tremendous research landscape for autonomous navigation and control using camera and LiDAR as its main exteroceptive sensors has already presented great successes [5], [7]. Next to the successes these camera and LiDAR methodologies have achieved, comes widespread adoption and a large community that has tested and improved these algorithms [8], [9], [10]. The use of these developed methodologies and algorithms on other sensing modalities would thus be desirable, as this would allow to easily extend the current sensing systems to increase the capabilities of the autonomous vehicle. However, the usage of existing methods developed for LiDAR usually does not generalise well on other senors, for example, sonar. In our mission to provide easy adoption and deployment of the 3D in-air ultrasonic sensor (eRTIS) we developed within our research group [11], we recently experimented with the prediction of LiDAR point cloud data based on the measurements of our eRTIS sensor using deep learning [12]. In this previous work, we demonstrated a convolutional stacked autoencoder, trained on eRTIS and LiDAR data that can accurately predict how a LiDAR sensor would have perceived the environment based on the measurements of the eRTIS sensor. In this paper, we extend this work towards the framework presented in figure 1 where the predictions are used in a LiDAR-based SLAM method. Section II provides background for the proposed approach introduced in section III. Finally, section IV discusses the research results and introduces potential future directions.

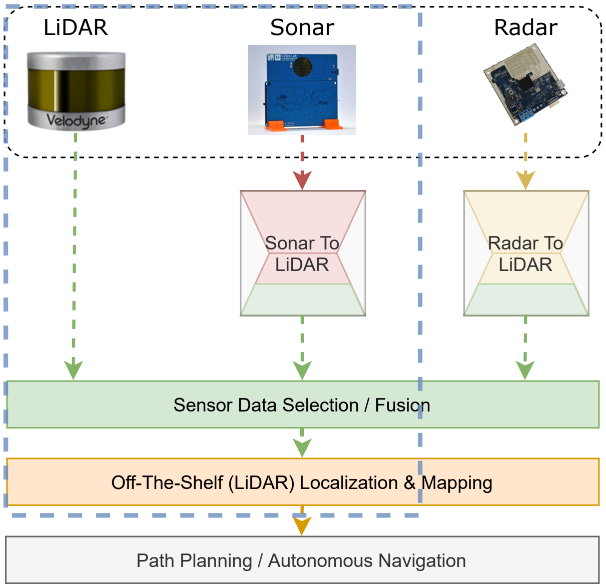
LiDAR
Sonar
Radar
Velodyne
Sonar To LiDAR
Radar To LiDAR
Sensor Data Selection / Fusion
Off-The-Shelf (LiDAR) Localization & Mapping
Path Planning / Autonomous Navigation

Fig. 1. Overview of the proposed method, which aims at abstracting the sensors used from further processing or the navigation stack. By converting the measurements from a multimodal sensing system into the desired representation, the navigation stack can be static while selecting the best possible measurement, allowing easy integration of novel sensors into existing systems. The work presented in this paper comprises the blue-dashed marked section.

II. Background

The following subsections briefly introduce S2L-Net (read: Sonar to LiDAR net) and similar state-of-the-art research that

* IDLab - Faculty of Applied Engineering, University of Antwerp - imec, Sint-Pietersvliet 7, 2000 Antwerp, Belgium

† Cosys-Lab - Faculty of Applied Engineering, University of Antwerp, 2000 Antwerp, Belgium

‡ Flanders Make Strategic Research Centre, Lommel, Belgium

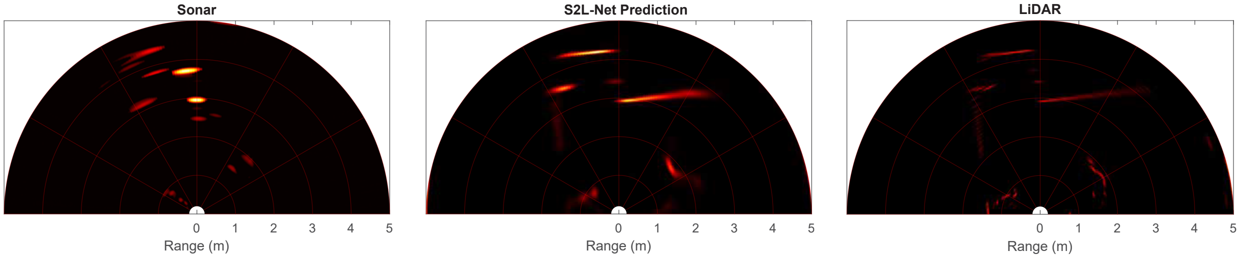

Fig. 2. Sample prediction of S2L-Net. The left image shows a real sonar energyscape measurement, made in an indoor office environment, used as the input of the model. The middle and right plot present the corresponding S2L-Net LiDAR prediction and ground truth, respectively.

also shows the benefit of modality prediction in new and existing sensing systems.

A. LiDAR point cloud prediction

In [11] our research group (CoSys-Lab) introduced a novel 3D in-air sonar sensor (eRTIS), which is capable of generating accurate 3D images of the environment using ultrasound. Recently, we proposed a method to transform the sparse ultrasonic measurements of the eRTIS into a dense point cloud representation, just as a LiDAR sensor would have perceived the environment. This can be seen as an ill-posed inverse problem, which has likely not a unique solution, therefore, we opted for a data-driven approach to obtain the prediction model. Our model, a convolutional stacked autoencoder, was trained on both simulated and real samples and achieved high accuracy on previously unseen data. Figure 2 depicts a sample prediction of S2L-Net on real sonar measurements of an indoor (office) environment. For a detailed discussion on the prediction results, e.g. a discussion on the generalisation of our model we redirect the reader to [12].

B. Different modalities in mapping and navigation

Many interesting research projects can be found that advocate for the use of different sensing modalities instead of the conventional camera and LiDAR approaches. However, most of these proposed methods do not work on the existing systems (e.g. LiDAR SLAM), making integration of the sensors difficult. A similar approach to this research is Milimap presented in [13]. It proposes a GAN model to generate or predict dense LiDAR-like occupancy grid maps with semantic annotations based on sparse mmWave radar measurements. The proposed approach is, however, limited to generating occupancy grid maps, whereas the system proposed by us provides point cloud data that can be used in any algorithm that accepts it.

III. S2L-SLAM

The goal of this research project was is to evaluate the usage of the prediction results of our S2L-Net in off-the-shelf algorithms with slight or no changes to the original algorithms. For this manuscript, we limit the scope of the discussion to LiDAR SLAM as this already provides an interesting use case and benchmark for the accuracy and usefulness of the predictions.

A. S2L-Net enhancements

We updated S2L-Net to optimise the output range-image for the conversion to a point cloud representation and to increase performance (accuracy, memory requirement and execution time). The first, most drastic change, was to the architecture of the model. We reduced the number of convolution filters, while we increased the number of convolution layers. The reasoning behind this comes from the limited perceptive field a convolution kernel has in each layer. For information to be exchanged between far-away data points in the sonar measurement it is, therefore, better to increase the network depth, while keeping the convolution filters smaller, resulting in a network size benefit. We also updated the input of the model to only include one sonar measurement since in the initial version we used five consecutive sonar measurements at the input. During the evaluation we concluded that one sonar measurement at the input provides sufficient accuracy while the network size can be greatly decreased, providing a significant memory and calculation time benefit. The second change was implemented during the training of the model. Our initial version of S2L-Net was trained using only Mean-Square-Error (MSE) Loss between the prediction and the LiDAR range image. As we convert the output to point cloud data by taking the index of the maximum value for each angle, we updated our loss function to optimise the output for this conversion. This can be seen as a classification task for each angle, where we have 500 classes for a range from 0 to 5 meters with 1 cm resolution. In short, we added a cross-entropy loss for every angle in the output image and the new loss function becomes:

$$Loss = MSE + Cross\ Entropy(for\ each\ angle) \quad (1)$$

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \frac{1}{N}\sum_{i=1}^{N}(-\frac{1}{M}\sum_{j=1}^{M}(y_i \log \hat{y}_i)) \quad (2)$$

Where N is the batch size and M is the number of angles, which in our example is equal to 180, $Y\ and\ \hat{Y}$ represent the ground truth and predicted range-images respectively, $y\ and\ \hat{y}$ represent the ground truth and predicted range class for each angle respectively. A Softmax layer at the output of the network is also added that operates on the angles (x-axis) of the output image. Mixing both losses (with equal weights) enables the model to calculate and recognise out-of-bounds (>5 meters) point predictions.

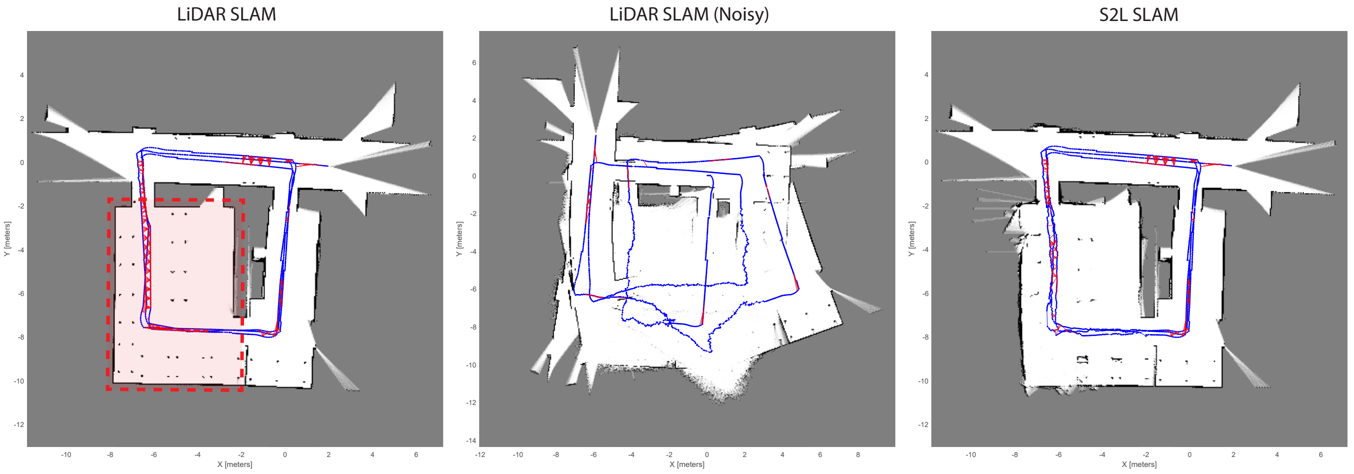

Fig. 3. Comparison between LiDAR SLAM and S2L-SLAM in a noisy environment on simulated LiDAR scans and sonar measurements. All sensor measurements are made within our MATLAB simulation environment. In the red marked box on the left plot we simulate a densely smoke-filled room. For each case, we plot the corresponding occupancy map and pose-graph obtained by the LiDAR SLAM algorithm found in MATLAB. The left plot presents the LiDAR ground-truth result, with no simulated noise. The middle plot shows the result when we keep using the noisy LiDAR scans, the right plot presents the result when we switch to S2L-Net scans in the noisy room. Note that no modifications to the SLAM method were made.

B. S2L predictions in SLAM

For usage of the S2L-Net predictions in off-the-shelf LiDAR-based algorithms, we first convert the output range-image to a point cloud representation. This conversion is done for each direction, by taking the Softmax of its corresponding range pixels (column). The index of the maximum value will represent the distance for that angle. This is essentially the argmax of the corresponding range column in the image for each direction. By applying Softmax over the range pixels, separately for every angle at the end of the network, we can interpret the pixels as confidence values for the corresponding range ”class”. Thresholding this confidence value determines whether the derived point falls in a specified range (determined by the sonar sensor). This thresholding works because of the MSE loss during training, when no class is available the confidence (pixel) values will be lowered. This derived point cloud representation is then used in the LiDAR-based SLAM algorithm. As figure 1 provides, a data selection or fusion layer can be added between the sensors and the navigation stack. For the LiDAR-based SLAM algorithm used in this paper, we opted for the LiDAR SLAM method readily available in MATLAB, which is based on the work in [14].

IV. Results and discussion

To demonstrate and verify our modality prediction framework, we used our simulation environment and an off-the-shelf LiDAR SLAM method. We simulated a robot within an office environment and navigated the same trajectory multiple times. The robot is equipped with a sonar and LiDAR sensor, both oriented similarly and with a field of view of 180 degrees. One of the rooms on the trajectory is densely filled with smoke, affecting the LiDAR measurements. The performance of the LiDAR SLAM algorithm is examined in this simulated noisy environment by switching to S2L-Net scans in the noisy room and comparing the result to using the noisy LiDAR scan. The results of this test are depicted in figure 3. This example aims to demonstrate the separation of the sensors used from the SLAM algorithm, which we can achieve by transforming different modal data into the correct modality. It is clear that in the smoke-filled room the position of the robot becomes uncertain when the noisy LiDAR measurements are used. The S2L-SLAM result in figure 3 shows when we switch to the predicted LiDAR measurements. Note that we do not change the processing stack, i.e. the SLAM algorithm, only the source of the point cloud data is switched when the LiDAR measurements are not suitable. For this S2L-SLAM result, approximately 1000 LiDAR scans and 500 predicted LiDAR scans were used. This clearly shows the benefit of using multimodal sensing setups as the S2L-SLAM is still able to accurately calculate an occupancy map and localise within it. This paper focuses on the usage of our modality prediction method to increase performance and extend the capabilities of existing systems. While the S2L-Net predictions can be used in all algorithms accepting LiDAR scans, we wanted to underline the benefit of using multimodal sensing setups, and showcase that we can dynamically select the best sensing modality without switching the processing stack. Based on the results presented in this paper, we conclude that our modality prediction framework certainly has the potential to succeed in its aim to separate the sensors used from further processing and allow for dynamic sensor selection or fusion. We believe this approach allows for easy integration of different modalities in existing systems when a suitable conversion model can be obtained or trained. In an extension upon this manuscript, the performance will be further explored and quantified on real measurements and radar sensors.

Acknowledgment

This work was supported by the Flemish Government (AI Research Program).

References

[1] A. Gilchrist, “Introducing Industry 4.0. In: Industry 4.0,” in *Industry 4.0*. Berkeley, CA: Apress, 2016, pp. 195–215. [Online]. Available: http://link.springer.com/10.1007/978-1-4842-2047-4

[2] T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade, and G. P. Li, “Predictive maintenance in the Industry 4.0: A systematic literature review,” *Computers and Industrial Engineering*, vol. 150, p. 106889, dec 2020.

[3] A. M. Farid, M. Alshareef, P. S. Badhesha, C. Boccaletti, N. A. A. Cacho, C. I. Carlier, A. Corriveau, I. Khayal, B. Liner, J. S. Martins, F. Rahimi, R. Rossett, W. C. Schoonenberg, A. Stillwell, and Y. Wang, “Smart City Drivers and Challenges in Urban-Mobility, Health-Care, and Interdependent Infrastructure Systems,” *IEEE Potentials*, vol. 40, no. 1, pp. 11–16, jan 2021.

[4] C. Kyrkou, S. Timotheou, P. Kolios, T. Theocharides, and C. Panayiotou, “Drones: Augmenting our quality of life,” *IEEE Potentials*, vol. 38, no. 1, pp. 30–36, jan 2019.

[5] D. Balemans, S. Vanneste, J. de Hoog, S. Mercelis, and P. Hellinckx, “LiDAR and Camera Sensor Fusion for 2D and 3D Object Detection,” in *Lecture Notes in Networks and Systems*. Springer, nov 2020, vol. 96, pp. 798–807. [Online]. Available: http://link.springer.com/10.1007/978-3-030-33509-0_75

[6] C. Chen, B. Wang, C. X. Lu, N. Trigoni, and A. Markham, “A Survey on Deep Learning for Localization and Mapping: Towards the Age of Spatial Machine Intelligence,” *arXiv*, jun 2020. [Online]. Available: http://arxiv.org/abs/2006.12567

[7] C. Debeunne and D. Vivet, “A Review of Visual-LiDAR Fusion based Simultaneous Localization and Mapping,” *Sensors*, vol. 20, no. 7, p. 2068, apr 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/7/2068

[8] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[9] G. Grisetti, C. Stachniss, and W. Burgard, “Improved techniques for grid mapping with Rao-Blackwellized particle filters,” *IEEE Transactions on Robotics*, vol. 23, no. 1, pp. 34–46, feb 2007.

[10] S. Kohlbrecher, J. Meyer, T. Graber, K. Petersen, U. Klingauf, and O. Von Stryk, “Hector open source modules for autonomous mapping and navigation with rescue robots,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8371 LNAI. Springer Verlag, 2014, pp. 624–631.

[11] R. Kerstens, D. Laurijssen, and J. Steckel, “eRTIS: A Fully Embedded Real Time 3D Imaging Sonar Sensor for Robotic Applications,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, may 2019, pp. 1438–1443. [Online]. Available: https://ieeexplore.ieee.org/document/8794419/

[12] N. Balemans, P. Hellinckx, and J. Steckel, “Predicting LiDAR Data from Sonar Images,” *IEEE Access*, vol. 9, pp. 57 897 – 57 906, apr 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9400352/

[13] C. X. Lu, S. Rosa, P. Zhao, B. Wang, C. Chen, J. A. Stankovic, N. Trigoni, and A. Markham, “See Through Smoke: Robust Indoor Mapping with Low-cost mmWave Radar,” in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*. New York, NY, USA: ACM, jun 2020, pp. 14–27. [Online]. Available: http://arxiv.org/abs/1911.00398 https://dl.acm.org/doi/10.1145/3386901.3388945

[14] W. Hess, D. Kohler, H. Rapp, and D. Andor, “Real-time loop closure in 2D LIDAR SLAM,” in *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2016-June. Institute of Electrical and Electronics Engineers Inc., jun 2016, pp. 1271–1278.