2022 NDIA GROUND VEHICLE SYSTEMS ENGINEERING AND TECHNOLOGY SYMPOSIUM
Autonomy, Artificial Intelligence & Robotics (AAIR) TECHNICAL SESSION
AUGUST 16-18, 2022 - NOVI, MICHIGAN

LIDAR SEMANTIC SEGMENTATION WITH A MULTI-RESERVOIR ECHO STATE NETWORK FOR OFF-ROAD TERRAIN PERCEPTION

S. Gardner[1], M. R. Haider[1], P. Fiorini[3], S. Misko[1], J. Smereka[2], P. Jayakumar[2], D. Gorsich[2], L. Moradi[1], and V. Vantsevich[1]
[1]The University of Alabama Birmingham, Birmingham, AL
[2]Ground Vehicle Systems Center (GVSC), Warren, MI
[3]Naisense

ABSTRACT

Autonomous vehicle perception has been widely explored using camera images but is limited with respect to LiDAR point cloud processing. Furthermore, focus is primarily on well-regulated environments, obviating a need for an algorithm that can contextualize dynamic and complex conditions through 3D point cloud representation. In this report, an Echo State Network for LiDAR signal processing is introduced and evaluated for its ability to perform semantic segmentation on unregulated terrains, using the RELLIS-3D open-source dataset. The L-ESN contains 16 parallel reservoirs with point cloud processing time of 1.9 seconds and 83.1% classification rate of 4 classes defining terrain trafficability, with no prior feature extraction or normalization, and a training time of 31 minutes. A 2D cost map is generated from the segmented point cloud for integration as a perception node plug-in to system-level navigation architectures.

Citation: S. Gardner, M. R. Haider, P. Fiorini, S. Misko, J. Smereka, P. Jayakumar, D. Gorsich, L. Moradi, and V. Vantsevich, “Lidar Semantic Segmentation with a Multi-reservoir Echo State Network for Off-road Terrain Perception”, In Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium (GVSETS), NDIA, Novi, MI, Aug. 16-18, 2022.

1. INTRODUCTION

Autonomous perception for vehicles targets the simplification of environmental information such that navigation response may receive occupancy grids or cost maps to perform autonomous maneuvering. This function remains a challenging active research domain for achieving true level-5 driverless cars [1]. Contextualizing the scene’s dynamic and chaotic conditions in a fast manner is critical for attaining targeted vehicle velocities safely and reliably.

The task of perception in the context of hostile off-road environments presents specific challenges. The environment is highly un-structured and changing, making differentiation between similar events a struggle. Other vehicles, people, and trafficable boundaries can be occluded in much more complex patterns than is the case with well-defined road traffic. Adverse elements can be purposefully or inadvertently concealed by taking advantage of the natural environment.

Beyond the challenges posed by the natural environment, there is a lack of pertinent datasets needed to meet the complete spectrum of scenarios. Most openly available datasets are built using urban

DISTRIBUTION A. Approved for public release; distribution unlimited. OPSEC#6583 settings, on-road operations, narrow ranges of anticipated events, and are recorded using the most common sensors, making training of customized sensor systems difficult.

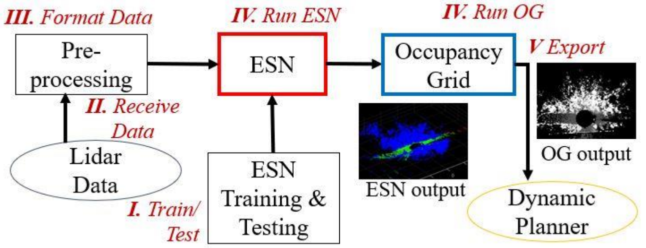

Figure 1. Perception Node Flow Diagram. The L-ESN and cost map are packaged within a framework that may be integrated into system-level architectures, communicating through ROS2

Contextualizing the scene's dynamic and chaotic conditions in a fast manner is the primary research motive of this work, critical for attaining targeted vehicle velocities. Autonomous perception can be broken into sub-functions: object detection for the identification and positioning of elements in the scene, region classification for the evaluation of terrain conditions surrounding the vehicle, and object tracking for tactical and defensive maneuvering. Each of the perception topics listed above are active research areas, notably in the context of autonomous driving, each with a large corpus of literature [2-4]. The region classification task is the focal topic of this research and can in first approximation be assimilated to a semantic segmentation. Semantic segmentation of an image or 3D point cloud consists of labelling each pixel/point of the image/frame according to a set of predefined categories.

Semantic segmentation is usually applied to classification of foreground objects while here the intended result is the segmentation of background objects. This background scene does not present well defined edges and can continuously evolve from drivable to unpassable. The performance of learning models must consequently be assessed in this context. Furthermore, phenomena leading to difficult visual conditions should also be factored in to ensure the algorithm's robustness and ultimately the vehicle's ability to actively navigate. These hurdles are not usually considered in academic research since they present non-ideal or unrealized events seen on regulated roadways. This is partly addressed in this research from the representation of tire tracks in mud, puddles of water on a trail, and ultimately the perceived levels of traversability seen in unregulated terrain camera datasets [5-6].

This work addresses these issues by using an Echo State Network (ESN) to rapidly classify the point cloud of input LiDAR frames. Integration with larger systems as a perception node is included by generating a 2D cost map commonly required by navigation systems. The flow diagram of the perception node is outlined in Figure 1. This report explores the metrics of the modified ESN for LiDAR signal processing (L-ESN) using an online benchmarked dataset that uses complex off-road LiDAR point clouds. Section 2 reviews the mathematics of the L-ESN and explores the parallel ESN architecture. Section 3 covers the benchmark dataset and L-ESN testing parameters. Section 4 shows the results of the trials. Section 5 explains the cost map and integration with autonomous vehicle systems. Section 6 contains a discussion, followed by a conclusion in Section 7.

2. LiDAR Processing with the ESN

The Echo State Network (ESN) has been a popular approach to time-series signal supervised learning since 2008 [7-9], but few have explored it for LiDAR processing to this point. A challenge is that LiDAR point clouds are high-dimensional signals that have variable dynamics according to the behavior of the sensor, how quickly regions of interest change between frames, and overall point cloud complexity. This report explores a scalable option that reduces the need for extensive preprocessing or long training times by taking advantage of the fundamental ESN architecture.

Applying one recurrent neural network (i.e. one reservoir) to process a 1-D signal is the most common approach when using ESNs, but is not viable for high-dimensional signals like point clouds due to high neuron requirements, which exponentially increases processing times (O^2). Thus, the common approach is to generate a small set of feature values per point that are fed through the ESN to generate the desired classification. The output point cloud is generated one point at a time, and substantial preprocessing is required before the ESN operates. Instead, numerous parallel reservoirs may be used to process sections of the input, with a large output matrix being trained to interpret the states of the reservoirs and give a classification or prediction, referred to as a task. Such a concept has been considered in literature [10], but not heavily for complex image processing. This approach avoids the problem of exponential training/processing time of a single central reservoir and distributes the computational load among many smaller reservoirs, allowing for substantially greater neuron-to-input ratios.

The basic Echo State Network architecture consists of an input signal or image represented as a vector. The input is multiplied by a random input weight matrix ($\mathbf{W}^{\mathbf{in}}$) and then passed through the reservoir with weights $\mathbf{W}^{\mathbf{res}}$. The reservoir is a recurrent neural network of standard leaky integrator neurons, which acts to transform the linear data into a high-dimensional, separable domain space. The neurons take on a value according to the input, collectively called the state vector, as defined by Equation (1). In the equation, **X** is the input data, **W** are the weight vectors, **S** are the collected states of the neurons in the reservoirs, and α is the learning rate. Ridge regression in Equation (2) or Moore-Penrose Pseudo-inverse is then used to generate an output weight vector that can interpret the reservoir states and give a classification. β is a regularization term to prevent overfitting, *I* is the identity matrix, and **Y** is the classified output point cloud. With the output weights calculated, the ESN simply needs an input to generate a classification according to Equation (3).

$$S = (1-\alpha)S + \alpha \tanh(W^{res}S + W^{in}X) \quad (1)$$

$$W^{OUT} = Y^{TARGET}S^{T}(SS^{T} + \beta I)^{-1} \quad (2)$$

$$Y^{TARGET} = W^{OUT}S \quad (3)$$

Literature shows that using multiple smaller parallel and/or series reservoirs have more improved network performances than a single large reservoir [11]. This concept is applied to the modified ESN by having multiple parallel reservoirs that split the input image into equal portions as visualized in Figure (2), with the total neuron count being number of parallel reservoirs times neurons per reservoir.

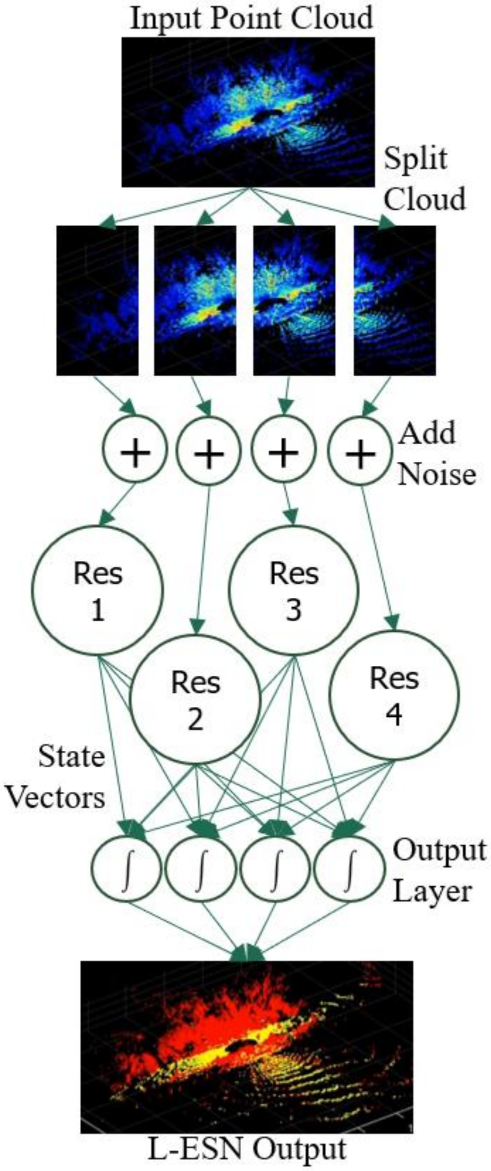

Figure 2. L-ESN Design Flow. The input point cloud is split, added with noise, and fed through the reservoirs. After they converge, the output weights are trained to classify the point cloud data.

The parallel reservoir approach classifies high volume inputs like the high-resolution point clouds of this work without exhibiting exponential training times associated with a single reservoir. Instead, training times increase linearly with the parallel reservoirs. The neuron states of each reservoir is concatenated into a single state vector, which can be defined as the input point cloud transformation into a linearly separable space that may be interpreted by the trained output weights. The parallel reservoir approach increases neuron-to-input ratio, needed to classify the LiDAR benchmark, as it has point clouds of greater size than most images.

A static input image independent of time can be represented as a time-series image for compatibility with the ESN by running the image through a standard Gaussian white noise filter multiple times to let the neurons in the reservoir converge and reach a classification. The added noise has been shown in many papers to improve classification results and is explained well in [12]. By training the algorithm to a noisier signal than the actual one, the features of a noise-free image are more identifiable to the model. Thus, the classification boundaries, generated by the reservoir and interpreted by the trained output weights, are more robust to noise and adaptable to subtle imperfections of the sensor system.

3. Testing Dataset and L-ESN Parameters

The tests of this report use a benchmark dataset that is then formatted to work with the L-ESN. The parameters of the ESN are essential for competitive performances. Training and testing were run on an Intel Core i7 vPRO at 2.3GHz. The setup of these components for the tests in this report is described in the remainder of this section.

3.1. Benchmark Dataset

The RELLIS-3D dataset was used for these tests [13]. The dataset contains color image sequences of complex off-road terrains such as trails, parks, and fields during ideal sunny conditions. They were taken using a camera mounted on a small mobile robot platform. The dataset has annotated images with 24 different classes. This report only uses three classes. Thus, the annotated LiDAR point clouds were reclassified according to four classes: (1) lowest resistance pathway, (2) moderate drivability, (3) low drivability, (4) non-drivable terrains. This report is among the first to use the ESN for LiDAR and non-binary output classifications.

3.2. L-ESN Testing Parameters

The L-ESN has many global parameters that define the system, with its performance depending strongly on what the values are initialized at before running the algorithm. As full network optimization is not within the scope of this report, a set of chosen parameters according to Table (1) have been used to generate the performance metrics of this report, and is based on running numerous trial runs and evaluating its output. The number of train/test samples in each epoch is split 80% train and 20% test from a randomly selected 120 images. The added white Gaussian noise has signal-to-noise ratio of 10 and the reservoirs will have 30 time-steps to converge upon a classification. These numbers are based on an understanding of the network dynamics and ability to perform quick evaluations from ultra-fast training times. A low spectral radius, input scaling factor, and learning rates are expected for signals exhibiting highly non-linearly separable data like discrete images.

Table 1. L-ESN Global Parameters. These settings produce the 4-classification results of this work.

L-ESN Parameters	Value
Train/Test Samples	120
Neurons	400
Reservoir Connectivity	10%
Input Scaling	1
Spectral Radius	0.008
Learning Rate	0.08
Time Steps	30

4. L-ESN Metrics and Analysis

With the global parameters defined, input signals pre-processed, and training complete, the L-ESN performances can be explored. As seen from the visualizations of Figure 3, the unregulated RELLIS-3D data is mapped to a corresponding 4-class semantic segmented point cloud that highly correlates with the actual annotated image. Point cloud classification error is calculated by subtracting the output and training image and then dividing by the total number of points. The classification rate for the parameters of Table 1 is 83.1%, with a point cloud processing time of 1.2 seconds, and training time of 31 minutes. Comparison to state-of-the-art algorithms using RELLIS-3D include the HRNet at 82.4% and GSCNN at 80.8%, both trained to segment LiDAR inputs into 19 classifications [14-15].

4.1. Architecture Optimization

While network optimization is outside the scope of this report, a grid search was performed on the number of parallel reservoirs and number of neurons per reservoir to understand how improved the classification is from higher point cloud-to-neuron ratios. With a total of 131,072 points, the maximum point-to-neuron ratios tested in this work is 82:1 which uses 1,600 total neurons. However, many more parallel reservoirs may be used to increase the neuron count. Performances generally improved for higher numbers of total neurons. For higher neuron counts, the training time increases. The extent of these trends is promising but require further investigation to understand how higher resolution and multi-sensor inputs affect the L-ESN architecture, error rates, and training time. Dissimilar datasets from the one used in this work will significantly alter the number of parallel reservoirs and number of neurons per reservoir that result in minimum error rates. Thus, when new data is introduced, the L- ESN can be quickly re-trained with a grid search to find the new optimal architecture. The duration/stability of this optimal point is proportional to the consistency of the environment. For wildly changing scenes such as single-cell thunderstorms or dust storms, the L-ESN performance is expected to dramatically decrease.

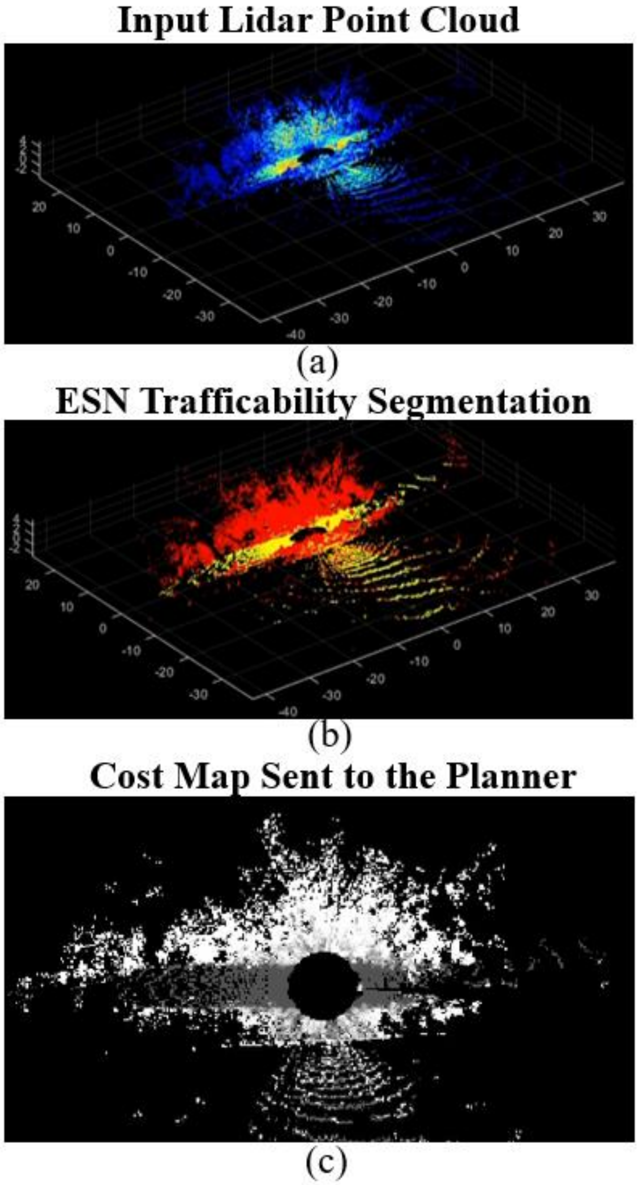

Figure 3. Example perception instance. (a) A LiDAR point cloud is sent to the perception algorithm. (b) The ESN receives the frame and classifies each pixel. (c) The segmented output is converted to a cost map.

5. Cost Map for Local Planning

Autonomous navigation typically requires occupancy information of the LiDAR data as contextualized by the perception algorithm. To complete the perception node, a cost map (or occupancy grid) is generated by filling a 500x500 2D grid with values [0, 0.33. 0.67, 1] representing highest, moderate, low, and zero trafficability, respectively. Figure 3c shows an example of the generated cost map that may be sent to the local navigation planner. White points represent the zero trafficability case while the darker points correspond to more drivable regions. An important distinction is noted that the background (any area with no associated points) is also black, making it visually seem like drivable region. However, only the defined points in the cost map are used for local mapping and planning. The L-ESN can be trained and imported as a pre-trained network along with the cost map. Communication is done via ROS2 and consists of initializing nodes for the L-ESN and cost map, such that they obtain LiDAR point clouds and process it continuously.

6. Discussion of Results and Future Work

The processing time per image is 1.9 seconds but should be within 60 milliseconds (i.e. 60 frames per second) for usage on moving vehicles, since the response time is critical for reacting to tire-soil mobility dynamics. Adaptive optimization of the architecture is a unique ability of the L-ESN since generally the network is established and unchanged during all training. There are clear trade-offs to using those CNN-based algorithms over the L-ESN, but for this application the L-ESN is a significantly more effective algorithm that can be scaled to contain a more robust ontology and adapt to new settings rapidly. Furthermore, the L-ESN can be adapted to different modalities that either perform faster training at the expense of error rates or vice versa. The biggest flaw of the L-ESN in this work is relatively high-performance instability when the finely tuned global parameters are changed, making optimization crucial to maintaining competitive error rates. The performance bounds when given significantly large reservoir sizes is to be explored in future works.

Inaccuracies of the manually labeled point clouds are mostly in the fine details, such as from grass, close-up objects, obscure objects, etc. Therefore, when the L-ESN detects objects that were not manually labeled, it implies that the real error rate of the L-ESN is lower by a small amount, as it slightly out-classified the labels at certain points.

In future works, this L-ESN approach will be explored for automatic feature extraction, faster processing speeds, automated hyper-parameter optimization, and usage with transfer learning techniques.

7. Conclusion

This work has shown an integration of a perception algorithm used in the novel way of processing LiDAR point clouds along with cost map generation for navigation planning and integration within autonomous vehicle systems. The algorithm has competitive classification rates, and low training requirements/processing times as desired for perception of off-road terrains. Test results show decreasing average point cloud classification error when increasing the number of parallel reservoirs and reservoir size. More investigation into the effects of reservoir size and the number of parallel reservoirs is in future works. Training takes minutes instead of hours, with as few as 120 training/testing samples, making it a promising approach to terrain mapping with unmanned autonomous vehicles.

8. Acknowledgements

This work is supported by Autonomous Vehicle Mobility Institute (AVMI) and University of Alabama at Birmingham (UAB), with special thanks to Ground Vehicle Support Center (GVSC) for overseeing this research.

REFERENCES

[1] Jebamikyous and Kashef, “Autonomous Vehicles Perception (AVP) Using Deep Learning: Modeling, Assessment, and Challenges”, ACM Computing Survey, vol. 10, 2022

[2] Boukerche and Hou, “Object Detection Using Deep Learning Methods in Traffic Scenarios”, ACM Computing Survey, vol. 54, 2021

[3] Lateef and Ruichek, “Survey on Semantic Segmentation using Deep Learning Techniques”, Neurocomputing, vol. 338, 201

[4] Cao and Bao, “A Survey On Image Semantic Segmentation Methods With Convolutional Neural Network”, Proceedings of CISCE, 2020

[5] Gardner, S., Haider, M. R., Smereka, J., Jayakumar, P., Gorsich, D., Moradi, L., & Vantsevich, V. (2021). Rapid High-Dimensional Semantic Segmentation with Echo State Networks. Ground Vehicle Systems Engineering Technology Symposium (GVSETS) - Autonomy, Artificial Intelligence, Robotics (AAIR).

[6] Gardner, S., Haider, M. R., Moradi, L., & Vantsevich, V. (2021). A Modified Echo State Network for Time Independent Image Classification. 64th IEEE International Midwest Symposium on Circuits and Systems (MWSCAS).

[7] E. Antonelo, B. Schrauwen, and D. Stroobandt, "Mobile robot control in the road sign problem using Reservoir Computing networks," in 2008 IEEE International Conference on Robotics and Automation, 2008, pp. 911-916.

[8] P. Yu, W. Jian-min, and P. Xi-yuan, "Traffic Prediction with Reservoir Computing for Mobile Networks," in 2009 Fifth International Conference on Natural Computation, 2009, vol. 2, pp. 464-468.

[9] F. Triefenbach, A. Jalalvand, and B. Schrauwen, "Phoneme Recognition with Large Hierarchical Reservoirs," in Advances in neural information processing systems (NIPS 2010), vol. 23Cambridge: MIT Press, 2010, pp. 2307-2315.

[10] X. Liu, M. Chen, C. Yin, and W. Saad, "Analysis of Memory Capacity for Deep Echo State Networks," presented at the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018.

[11] L. Manneschi, M. O. A. Ellis, G. Gigante, A. C. Lin, P. Del Giudice, and E. Vasilaki, "Exploiting Multiple Timescales in Hierarchical Echo State Networks," (in English), Frontiers in Applied Mathematics and Statistics, Original Research vol. 6, no. 76, 2021-February-17 2021.

[12] M. Lukosevicius, "A Practical Guide to Applying Echo State Networks," in Neural Networks: Tricks of the Trade, Reloaded, vol. 7700, G. Montavon, G. B. Orr, and K. R. Muller, Eds.: Springer, 2012, pp. 659–686.

[13] Jiang and al., “RELLIS-3D Dataset: Data, Benchmarks and Analysis”, arXiv, 2020

[14] Yuan, Y., Chen, X., Chen, X., & Wang, J. (2019). Segmentation transformer: Object-contextual representations for semantic segmentation. arXiv preprint arXiv:1909.11065.

[15] Takikawa, T., Acuna, D., Jampani, V., & Fidler, S. (2019). Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. ICCV.