Integrating Temporal Modeling into End-to-End Driving for Ultralight Collaborative Vehicles: Preliminary Results

Nosheen Abid[1,*], Paul Davidsson[1], Hannes Bergkvist[2] and Ivar Bergkvist[2]

[1]Malmö University, 205 06 Malmö, Sweden
[2]LEVTEK Sweden AB, Kosterögatan 15, 212 14 Malmö, Sweden

Abstract

This work presents a preliminary investigation in extending end-to-end autonomous driving systems with temporal modeling for collaborative ultralight robotic vehicles. As part of a larger FORMAS-funded project on human-robot collaboration, we explore integrating Long Short-Term Memory (LSTM) layers into the classic PilotNet architecture to enhance steering stability and responsiveness in real-world scenarios. Unlike prior work focused on simulation or large-scale driving datasets, our approach is tested on data collected from LEVTEK Sweden AB's ultralight robotic utility vehicles designed for sidewalk and indoor environments. Early results show that even short-term temporal context improves steering smoothness and reduces response lag. This initial study lays the groundwork for our ongoing FORMAS-funded research, focusing on integrating attention mechanisms, transformer-based models, imitation learning, and deployment in real-world collaborative environments. The long-term aim is to develop adaptive, explainable, and human-aware autonomy for lightweight robotic vehicles.

Keywords

Imitation Learning, Autonomous Driving, Ultralight Vehicles, Temopral Modeling

1. Introduction

Urban logistics, property management, and sidewalk delivery increasingly use collaborative robots. Most commercial and academic efforts rely on modular pipelines that split the driving task into perception, planning, and control [1]. Such designs are structured but often suffer from error propagation, where failure in perception can cascade into planning and control. In contrast, end-to-end learning maps sensory input directly to driving control using a single neural network. This reduces complexity, lowers latency, reduces the need for hand-crafted outputs, and improves robustness to distribution shifts and adversarial attacks [2, 3]. The concept began in the 1980s with ALVINN [4], showing a neural network could learn to steer from camera images [4]. Decades later, NVIDIA PilotNet [5] advanced it using a deep CNN to learn lane-following from human driving data. Since then, the field has expanded to include CNNs [6], reinforcement learning [7], and hybrid models enabling human-like driving behaviour [8].

However, early CNN-based end-to-end models rely on single-frame perception, lacking

SAIS2025: Swedish AI Society Workshop 2025, 16-17 June 2025, Halmstad, Sweden.
*Corresponding author.
nosheen.abid@mau.se (N. Abid); paul.davidsson@mau.se (P. Davidsson); hannes@levtek.io (H. Bergkvist)
0000-0002-5922-7889 (N. Abid); 0000-0003-0998-6585 (P. Davidsson)

temporal memory. In practice, driving is a sequential task. Humans predict upcoming turns, smooth out steering actions, and react over time. A purely feed-forward vision model may react too late or abruptly to changes This motivates integrating a recurrent neural network, such as a Long Short-Term Memory (LSTM), into the end-to-end model to capture temporal context. Temporal modeling enables the system to track how the environment and the vehicle's state change over time by remembering information from previous frames. This approach has been shown to perform well in diverse driving settings [9, 10], and remains widely adopted in advanced systems such as NEAT [11] and TransFuser [12]. Recent investigations, including ReasonNet [13], ST-P3 [14], and CIRL [15], demonstrate that memory modules greatly improves driving performance by allowing the network to anticipate and plan rather than merely react.

This work is conducted in the real-world context of collaborative autonomy. Unlike most self-driving systems that operate alone, we focus on ultralight robotic utility vehicles (from LEVTEK Sweden AB) that a person can ride or walk beside, switching easily between manual and autonomous modes. These vehicles are compact, low-speed, and designed for hybrid indoor-outdoor operation, making them ideal for working closely with people. For instance, a postal worker might sometimes drive the vehicle, and other times have it follow autonomously. Such human-robot collaborative driving demands not only end-to-end autonomy, but also the ability to follow human input and behave in a smooth, predictable way that people can trust.

This paper presents a preliminary study of a two-year FORMAS-funded project (2024-01126) "Human-robot collaborative learning for ultralight electric utility vehicles" with LEVTEK, PostNord, RETTA and MAU. This work extends PilotNet with an LSTM-based temporal module and trains it on real driving data collected from LEVTEK's prototype vehicles. The goal is to assess if memory-enhanced end-to-end models improve steering smoothness and responsiveness. The current results are limited in scope forming a foundation for the coming phases of the project. We cover methodology and initial findings in Section 2, discuss key insights and on current limitations in Section 3, and outline the future work and conclusion in Section 4.

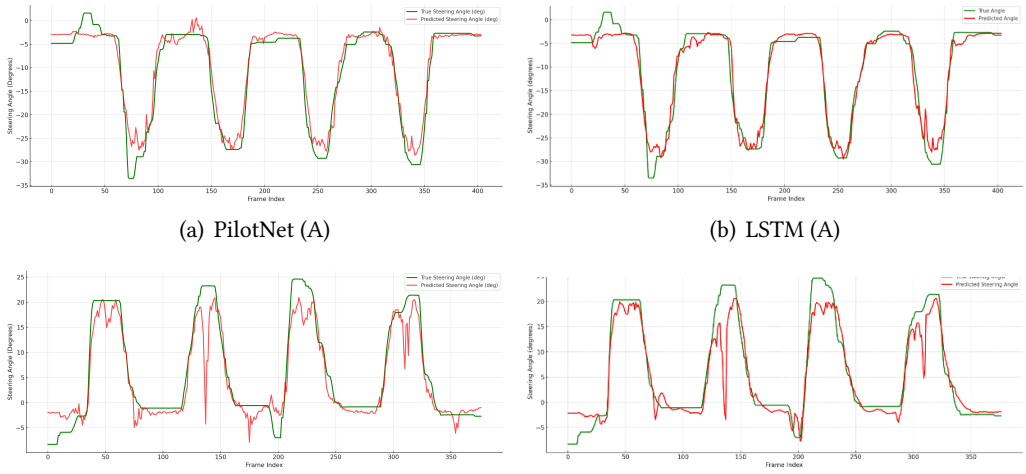
2. Preliminary Work

Data Collection: Experiments use LEVTEK's ultralight robotic utility vehicle, designed for both indoor and outdoor use. It's equipped with multiple cameras, lidar, and additional sensors, along with electronic controls for steering, braking, and throttle. This works uses the forward-facing monocular RGB camera to record several human driven sessions on a test course with indoor, hallway like areas. The dataset has 38 training (~50k frames) and 4 testing (~2.5k frames) sessions having time-synced frames and steering commands, covering straight paths, gentle curves, and a few sharp turns (up to ±30°). Although the dataset is small and has limited variations by deep learning standards, it still reflects the target environment of vehicle having low-speed, and cluttered surrounding. All images are downsampled to 200×66 and normalized, matching the original PilotNet setup. The low resolution keeps the model efficient and lightweight.

LSTM-Enhanced PilotNet Architecture: The PilotNet architecture is adapted to include temporal modeling by inserting an LSTM layer after the CNN feature extractor. The network takes a sequence of N consecutive frames (we used $N = 5$ in initial experiments) as input. Each frame goes through the same CNN to extract visual features. These features are then passed

to an LSTM with 512 hidden units, which helps the model remember what happened earlier, like when a turn started. The LSTM outputs its final hidden state after the Nth frame, which is then fed to PilotNet original Fully-Connected (FC) layers to predict the steering angle. The FC part of the network consists of a 1164-unit layer with ReLU, followed by 100 and 50-unit layers, and a final output layer. The entire model has about 250k trainable parameters, only slightly more than the original. Since we did not have a pretrained model for our domain, the model is trained from scratch with randomly initialized weights .

Training: Learning is framed as a regression problem, where the network predicts a steering angle to match the human driver using supervised imitation learning. The Huber loss (smooth L1 loss) is employed for its greater stability compared to mean squared error, particularly due to its robustness to outliers. To avoid overfitting, L2 regularization with a factor of 1e-3 is applied. Training runs for 25 epochs using the Adam optimizer (learning rate 1e-4) with a batch size of 128 sequences, each with $128 \times N$ frames.



(a) PilotNet (A) (b) LSTM (A)

(c) PilotNet (B) (d) LSTM (B)

Figure 1: Steering predictions comparison of PilotNet and LSTM-enhanced PilotNet on test cases.

Evaluation and Observations: At this early stage, the evaluation is open-loop, meaning it predicts steering on test sequences, rather than deploying it to drive the vehicle autonomously yet. We compare two versions: i) the original PilotNet, trained on single frames, and ii) LSTM-enhanced PilotNet , trained on 5-frame sequences.

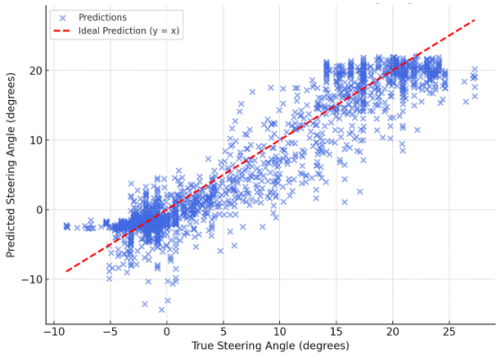


Figure 2: Scatter plot comparing predicted vs. true steering angles - LSTM-enhanced PilotNet.

The Figure 1 compares steering predictions from PilotNet and the LSTM-enhanced PilotNet across two test scenarios. In both (A) and (B), the LSTM model (right) tracks true steering angles (green line) more closely than PilotNet (left), especially around sharp turns and transitions. The LSTM reduces delay and overshooting, leading to smoother, more stable control, showing the benefit of adding temporal memory. In Figure 2,

the predicted steering angles of LSTM-enhanced PilotNet are compared with true values. Each blue dot is one prediction. The closer it is to the red line, the more accurate it is. Most points cluster near the red line. There is some scatter at higher angles, but overall the predictions follow true steering behavior. The model overall performs well, with strong Pearson correlation ($r = 0.95$) and R^2 score of 0.89, indicating it explains most variation in human steering. The average error remains low on straight roads (MAE $\approx 1.65°–2.28°$), while mild turns show slightly higher error (MAE $\approx 3.86°–4.57°$), indicating occasional under/over-correction. Interestingly, sharp turns (MAE $\approx 2.79°–3.61°$) are handled better than mild ones, suggesting the model detects turning intent, though fine control may still need improvement. These results demonstrate strong early-stage performance and highlight opportunities for further improvement.

3. Limitations and Insights

As this preliminary study is based on a few months of work, it has several limitations. So far, all evaluations have been conducted offline. The LSTM-enhanced model has not yet been deployed on the vehicle for real-time driving. Consequently this study lacks insight about how the model handles situations where small errors build up or whether it might cause unstable steering behavior when in full control. The training dataset is relatively small and biased, collected by a single driver on a fixed route. The performance of the model in unfamiliar situations remains untested. It is suspected that the model may have learned false correlations from the limited data. For example, associating a particular wall color with turn, due to repeating patterns in the training data. Despite using regularization techniques that gained reasonable performance on test data, the risk of overfitting remains. True generalization will only be assessed on more diverse routes. Another limitation is that the current end-to-end model has no explicit notion of obstacles or pedestrians. It is trained only to mimic steering within a mostly clear path, and does not produce outputs for braking or obstacle avoidance. PilotNet by design has no direct output for braking or obstacle detection. Although the LSTM helps reduce delay between perception and action, some physical lags remains when the actual vehicle actually moves. This highlights that simply learning the mapping from image to steering may not be enough if the vehicle has significant inertia or delay. A better approach could involve predictive control or feeding in extra information like the current speed to help the model predict future actions.

4. Future Directions and Conclusion

In conclusion, this early work shows that adding LSTM-based temporal modeling to an end-to-end driving model leads to smoother and more responsive control using real ultralight vehicle data. Future work will focus on enhancing performance and transparency through attention mechanisms, Transformer-based architectures for long-term memory, and using tools like saliency maps. A key direction is developmenting interactive learning, where human feedback actively shapes the model during training and deployment. This support the broader goal of building an autonomous system grounded in human-in-the-loop machine learning for adaptive, real-time human–machine collaboration. Expanding real-world testing and ongoing partnership between LEVTEK, are essential to bringing these advances to actual vehicle deployment.

References

[1] E. Yurtsever, J. Lambert, A. Carballo, K. Takeda, A survey of autonomous driving: Common practices and emerging technologies, IEEE access 8 (2020) 58443–58469.

[2] H. M. Eraqi, M. N. Moustafa, J. Honer, Dynamic conditional imitation learning for autonomous driving, IEEE Transactions on Intelligent Transportation Systems 23 (2022).

[3] P. S. Chib, P. Singh, Recent advancements in end-to-end autonomous driving using deep learning: A survey, IEEE Transactions on Intelligent Vehicles 9 (2023) 103–118.

[4] D. A. Pomerleau, ALVINN: An autonomous land vehicle in a neural network, Advances in neural information processing systems 1 (1988).

[5] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al., End to end learning for self-driving cars, arXiv preprint arXiv:1604.07316 (2016).

[6] F. Codevilla, M. Müller, A. López, V. Koltun, A. Dosovitskiy, End-to-end driving via conditional imitation learning, in: 2018 IEEE international conference on robotics and automation (ICRA), IEEE, 2018, pp. 4693–4700.

[7] Z. Zhu, H. Zhao, A survey of deep RL and IL for autonomous driving policy learning, IEEE Transactions on Intelligent Transportation Systems 23 (2021) 14043–14065.

[8] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, A. M. López, Multimodal end-to-end autonomous driving, IEEE Transactions on Intelligent Transportation Systems 23 (2020).

[9] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, A. Geiger, PlanT: Explainable Planning Transformers via Object-Level Representations, in: Conference on Robot Learning, PMLR, 2023, pp. 459–470.

[10] Y. Wang, D. Zhang, J. Wang, Z. Chen, Y. Li, Y. Wang, R. Xiong, Imitation learning of hierarchical driving model: From continuous intention to continuous trajectory, IEEE Robotics and Automation Letters 6 (2021) 2477–2484.

[11] K. Chitta, A. Prakash, A. Geiger, NEAT: NEural ATtention fields for end-to-end autonomous driving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15793–15803.

[12] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, A. Geiger, TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[13] H. Shao, L. Wang, R. Chen, S. L. Waslander, H. Li, Y. Liu, ReasonNet: End-to-end driving with temporal and global reasoning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 13723–13733.

[14] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, D. Tao, ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning, in: European Conference on Computer Vision, Springer, 2022, pp. 533–549.

[15] X. Liang, T. Wang, L. Yang, E. Xing, CIRL: Controllable Imitative Reinforcement Learning for vision-based self-driving, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 584–599.