End-to-End Joint Intention Estimation for Shared Control Personal Mobility Navigation

Kavindie Katuwandeniya, Jaime Valls Miro and Lakshitha Dantanarayana

Abstract— **Advancements in technology propose a future where systems work collaboratively sharing the same workspace as humans. Navigation is one such crucial aspect of daily life where collaborative technologies can offer major assistance. Ageing population dictates a likely increase in personal mobility devices (PMDs), whilst autonomous cars are bringing intelligent vehicles to the road today. However, in such scenarios the expected assistance can only be given if the device is aware of its user's intention, so that controls can be applied in a tightly collaborative manner. Moreover, they should be robust to different environments, users and mobile platforms. A user driven navigation framework is proposed in this work to complement end-to-end sensing-only solutions to estimate controls as joint intention from vehicle states and user inputs. The solution is proven to be an improvement over similar strategies that rely on exteroceptive data and omit inputs from the driving agent. Furthermore, the developed framework is proven capable of transferring the learning into different environments and mobility platforms using a small amount of training data. Data from the autonomous driving community (Udacity dataset) and other obtained in-house with an instrumented power wheelchair are given to demonstrate the validity of the proposed approach.**

I. Introduction

The discipline of Human Robot Interaction (HRI) is concerned with the understanding and shaping of the interactions between human and robotic agents [1]. Human Robot Collaboration (HRC) is generally regarded as the subset of HRI where the aim is to achieve a common goal through tight, shared interactions between human and machine [2]. It is an interdisciplinary research field, sitting at the intersection of classical robotics, cognitive sciences, and psychology. This shared control framework is often referred to as autonomy or arbitration [3], and has a significant constituent design impact in the overall HRI problem [1]. The authors of [3] have proposed a taxonomy on levels of autonomy based on the level of human-robot intervention in the sense, plan, and act phases, defining autonomy as the extent that a robot can sense, plan and act to reach a goal without an external control. The level of autonomy required in a situation is however highly specific to both application and user, and can extend across the full spectrum between fully manual and fully autonomous control. Recent examples of these can be found in the autonomous driving and driver-assistance systems, which have been active ares of research for the past two decades, to varying degrees of commercial success.

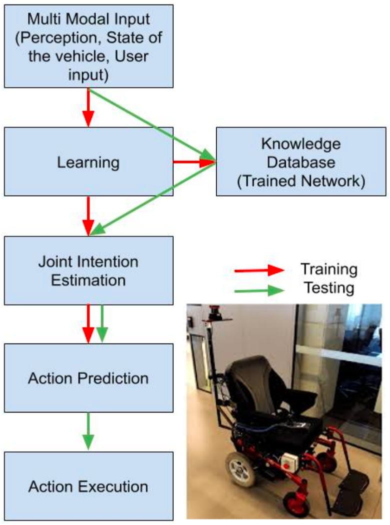

Multi Modal Input (Perception, State of the vehicle, User input)
Learning
Knowledge Database (Trained Network)
Joint Intention Estimation
Training
Testing
Action Prediction
Action Execution

Fig. 1: The HRC learning framework and the instrumented power wheelchair used for testing.

In the proposed work, examples from these fields are employed, and learnings are transferred to the personal mobility space to depict the paramount role that HRC plays in generating safe navigational paths that conform to the driver's intentions and expectations, and those around them. With the advent of increasingly advanced PMDs, be that to mitigate traffic congestion and pollution through reduction of road vehicle usage in the "first and last mile" of city travel [4], or as power mobility aids to promote independence and self-esteem for the differently-abled and frail users in an ever increasing globally aging population [5], [6], there is a strong motivation to develop user intention estimations and collaborative strategies suitable to integrate seamlessly with vehicle controls.

User intention can be interpreted in numerous ways and is not always straightforward to estimate. User intention has been formally defined as the combination of states of different subsystems in the body, namely: i) central nervous system (CNS), ii) peripheral nervous system (PNS), iii) musculoskeletal system, and iv) the controls generated for environmental interactions [7]. In this work the focus is on the latter, with the goal of HRC maneuvering of mobility devices, such as the motorised wheelchair shown in Fig. 1. In this regard, user intention can be broadly confined to the immediate controls of the vehicle, longer-horizon trajectory control, or the actual goal destination (e.g. map-based). The aim hereby is the consideration of immediate shared controls - steering and speed of the mobility device - at the next time instance.

In typical HRC research scenarios, a framework infers the agent's input controls based on the sensory and/or vehicle state inputs, to be then compared with the inferred user intention (e.g. steering angles from on-board car images [8]). In others, actual controls are exerted from the joint intention deduced from both agents, whereby the autonomous agent has been equipped with some autonomous map-based path planning capabilities. This is often based on some pre-defined weighted function (e.g. "efficiency" [9]), which not only requires tuning depending on the scenario, but also offers limited applicability to larger, or map-less situations.

In this work a learning network is proposed to provide controls (immediate speed and steering) for shared navigation based on joint intention estimation, and prove that the developed framework is capable of transferring into different environments (indoor, outdoors), and platforms (cars and PMDs). The framework is extendable to any situation where tight human-robot driving collaboration is assumed, e.g. warehouse management with instrumented forklifts, museum tours or grocery shopping with mobility scooters, or added independence for the differently-abled with PMDs.

The paper is organized as follows: Section II summarises relevant work from the literature. Section III describes the proposed framework. Section IV describes the experiments carried out and Section V summarizes the result. Discussion, future work and conclusion are given in Section VI.

II. Related Work

There are two widely used tools for changing roles in the shared control space: i. machine learning, and ii. performance metric based [7].

A large volume of work stems from user intention-aware for the differently-abled, largely aligned with smart wheelchairs [10], which are invariably driven by performance metric based controls. Early work like NavChair [11] have different modes of navigation which are triggered by pre-defined values such as stimulus response. [12] proposed a shared control system which gives necessary assistance based on the user ability. In more recent work, the decision of autonomy in control is based on a dynamic value based on risk of collision and the risk of disturbance to other humans [13].

The main drawback in a performance metric based role allocation is that the switching mechanism becomes an autonomous entity, which tend to leave users confused and somewhat apprehensive about its behaviour. On the other hand, when triggered by a human operator, the framework lacks the ability to identify user's intention and act accordingly. Additionally, for the case of differently-abled users, they learn to rely more on the framework and lose residual capabilities [12].

On the other hand, the advent of research in the space of autonomous cars have leaned strongly on machine learning architectures for role changing. During the past two decades, the advancements in hardware (sensors and actuators) and software (advanced algorithms, deep neural networks) have led to the rapid developments in the field. Steering control is a crucial aspect of autonomous navigation. Much work has been done in the field of steering angle prediction and companies like Udacity [14] have even released challenges based on this to make the research community more engaged. [15] provides a literature review in this regard.

A steering angle prediction framework has been recently proposed in [8] where the authors address issues overlooked in previous research work by encoding spatiotemporal features at different layers. Their results rank best when predicting the current steering angle in comparison to the existing methods. Their work is predicated on using only images as input. [16] integrates speed, wheel angle and torque in addition to the images and prove that the: i. addition of extra information, ii. having residual aggregation, and iii. having ConvLSTM, improve the accuracy in steering angle prediction. The major drawback in this work is that even though they integrate additional information, this information is predicted from the system itself. Neither the real user input nor the vehicle state is considered in their work. In contrast, our proposed framework considers the user in the prediction and control loop.

The network proposed by [17] integrates images and past 10 speeds to predict the immediate steering and speed. Their framework too does not consider the user input. [18] and [19] consider the user input in their developed frameworks. [19] uses three modes: i. direct, ii. follow, and iii. fugitive mode as the modal information; while in [18], the user commands are a limited vocabulary consisting of phrases like "keep straight", "turn right". The major difference our work has from theirs is that their user commands (intentions) are given explicitly while in ours it is inferred.

In the proposed framework there is no explicit role change, advocating instead for shared controls predicted based on joint intention estimates implicitly by a machine learning paradigm, with perceptual information from cameras, vehicle state, and user commands (joystick or steering wheel angles) used as inputs.

III. Proposed Framework

This work aims to develop a shared navigation framework for action planning and execution based on inferred joint intention. At one end of the spectrum of systems developed to enable autonomous driving are methodologies behaving in an end-to-end manner, where sensor inputs are directly mapped to actions using various function approximators [18]. Deep Neural Networks have become widely used and have been proven successful non-linear function estimators, and have thus been employed in this work as the *learning* module for the proposed joint intention estimation HRC framework. An overview of the framework in depicted in Fig. 1, whereby the joint intention is inferred based on muti-modal inputs, namely past and current user inputs, vehicle states and environmental observations. For the purpose of this research steering angle and speed of the mobility device at time $t+1$ (predicted at t) are considered as the joint intention. Three frameworks are considered to achieve this purpose.

1) Network Architecture I: Image Only Input: The inputs of this network are the past n ($t-n+1$ to t) images and it predicts steering angle at $t+1$. This was implemented based on the network proposed by [8] and the details of the implementation are given in Section IV-.1.

2) Network Architecture II: UDM (User Driven Multi-Modal): The inputs consist of past n ($t-n+1$ to t) images, vehicle speed and user steering angle and the output is the steering at $t+1$. The linear vehicle speed is the vehicle's

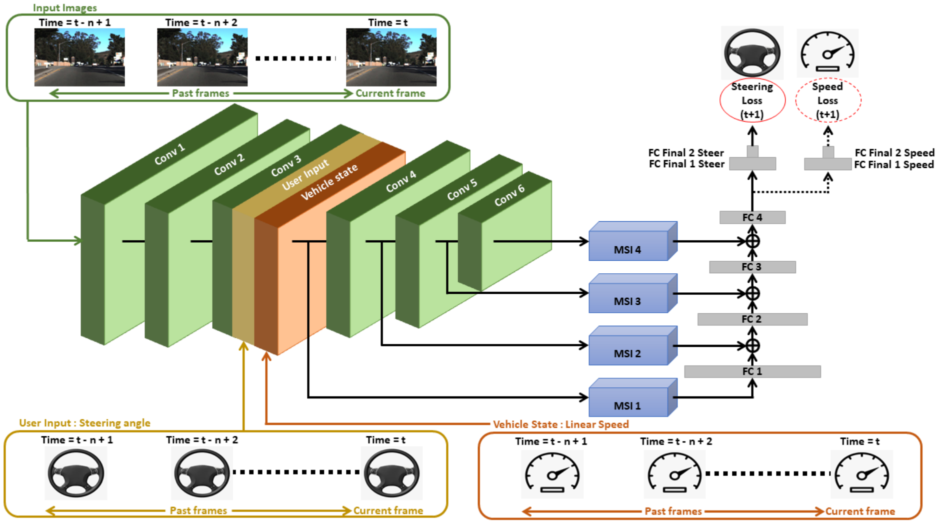


Fig. 2: Proposed Framework. UDMT and UDM (with and without the dashed black line branch in the output respectively).

state input and the steering angle is the user input. The best place to add the user input and the vehicle state to the Implementation I network is described in Section IV-.2.

3) Network Architecture III: UDMT (User Driven Multi-Modal Multi-Task): As before, input to this framework are the past n ($t - n + 1$ to t) images, vehicle speed and user steering angle, but it now predicts steering plus speed at $t+1$. The rationale is that in order to control the vehicle, speed may be as important as steering, so the network is setup to also predict speed. The modified framework is described in Section IV-.3.

A. Intuition Behind the Proposed Framework

The optimisation of the set of parameters θ that define the deep neural network trained with images only can be described by

$$\theta^* = \arg\min_\theta \sum_t \text{MSE}(F_\theta^{\text{Image}}(o_{n,t}), [\phi_{1,t+1}]) \tag{1}$$

where an input sample at time t consists of a history of the past n observation images $o_{n,t}$. The network tries to find the set of parameters of the function approximator F_θ^{Image} which minimizes the mean squared error (MSE) between the predicted and the actual steering angle at $t + 1$ (ϕ_{t+1}).

This has the implicit assumption that the steering angle is a function of only the images and the parameters of the network. However, it is evident that this is not the case. Consider a case where the user is at a four-way junction. The steering angle at this point, can have three different values which are equally probable. The image only network will arbitrarily choose a direction based on the learning which might not be the intended direction of the user. Such a network, even if it avoids collision and take human-like turns at intersections, ignores the desire of the rider in the vehicle.

The proposed framework attempts to correct this drawback. If the user intention can be better inferred from the data provided, the ambiguity in steering of the car is mitigated. The framework is built on the assumption that the joint intention of both user and agent can be estimated more accurately by including past speed and steering in addition to the observation images, as given by function E,

$$\text{Joint Intention}_{t+1} = E(o_{n,t}, \phi_{n,t}, v_{n,t}) \tag{2}$$

In this equation, the user and agent intentions can be thought of as latent variables. From the joint intention, shared-control actions are estimated through function G:

$$[\phi_{1,t+1}, v_{1,t+1}] = G(\text{Joint Intention}_{t+1}) \tag{3}$$

The joint intention is mapped to the immediate future steering angle and speed (control actions). As our model is trained end-to-end, the learned function F_θ combines the joint intention estimation E and shared-control prediction G functions. Thus, equations (2) and (3) can be reduced to F_θ.

$$F_\theta = G \circ E \tag{4}$$

The final objective function of the network is written as

$$\theta^* = \arg\min_\theta \sum_t \text{MSE}(F_\theta^{\text{UDMT}}(o_{n,t}, \phi_{n,t}, v_{n,t}), [\phi_{1,t+1}, v_{1,t+1}]) \tag{5}$$

where θ is the set of parameters of the trained network. Here, an input sample at time t consists of a history of the past n observation images $o_{n,t}$, steering angles $\phi_{n,t}$, and vehicle speeds $v_{n,t}$, from $t - n + 1$ to t. The network tries to find the set of parameters of the function approximator F_θ^{UDMT} which minimizes the mean squared error (MSE) between the predicted and the actual ϕ_{t+1} and v_{t+1}.

During the testing phase, the same input modalities are used and inference is done based on the knowledge database created from learning as shown in Fig. 1.

IV. Implementation

1) Network I: Input Images Only: It was decided to use the network proposed in [8] to detect features as authors claim their network to outperform other existing networks in terms of network accuracy. In order to replicate the above network, kernel sizes of 3 and 5 were used with different

TABLE I: Details of the network

Layer Name	Kernel Size	Stride Size	Padding
Encoder			
Conv 1, Conv 2, Conv 3	5 x 5	2	none
Conv 4, Conv 5, Conv 6	3 x 3	2	none
Multi-scale Spatiotemporal Integration (MSI) Module (* = 1,2,3,4)			
Conv *-1	3 x 3	1	1, zeros
Conv-LSTM *	3 x 3	1	1, zeros
Conv *-3	3 x 3	2	none
Conv *-4	3 x 3	1	1, zeros
FC *	-	-	-
Predictor			
FC final 1 , FC final 2	-	-	-

stride sizes and padding which results in the stated output sizes of the layers. While there are no specific details about batch normalization (BN) being used, it is standard practice and was incorporated after every convolution layer [20]. *Leaky-ReLU* activation function is used after all layers except the last fully connected layer, which is linear. Full details of the network architecture are collected in Table I.

The convolution layers at the end of each Multi-scale Spatiotemporal Integration (MSI) module was flattened and concatenated with the previous FC layer.

The input image sequence length n was chosen to be 10, same as in the [8]. The authors of the paper have trained different networks configurations with steering angles at $t+k$, where k = 1 , ... , 4 as future auxiliary labels to predict the steering at t. In the proposed network, auxiliary future labels are not used for training which is equivalent to their baseline MSINet. Furthermore, the proposed framework predicts the next steering (at $t+1$) with current and past images ($t-n+1$ to t) while they predict steering at t with the same input images.

2) Network II: UDM: In this implementation the state of the vehicle and the user input is integrated into the image only network IV-.1. As the state of the vehicle, the current and past ($n = 10$) linear speeds are considered. The most crucial aspect of the framework is the user input, which is the proposed way of including the user in the control loop. As the user input, the current and past ($n = 10$) steering angles are considered.

Since the proposed framework uses vehicle state and the user input in addition to images, the best way to integrate the new inputs has to be selected. According to the authors of [18], this can be done either using deep autoencoders or using concatenation. It was decided to use concatenation, and consider deep autoencoders for future work.

Most work in this field has the vehicle state integrated after the images' dimensionality is reduced [16], [17] (eg: The 1D feature vectors are appended with the vehicle state and fed into LSTMs). This is not an optimal way of doing it, since the influence speed has on the feature detection becomes low at later layers of the deep network.

The best place to concatenate the new inputs was selected to be at layer 3 (Conv 3) before the branching of the encoder into MSI modules. This way the first few layers are not effected by the additional inputs in terms of the features they select (more generic image features are selected) and the later features are influenced by the additional inputs. By concatenating before the LSTM, temporal relation is still preserved. The speed and the user steering of past n frames, are reshaped to match the size of the Conv 3. The above is achieved by tiling a shape of the desired shape (57×77) and concatenating the layers as depicted in Fig. 2.

(a) Image from Bag 4. (b) Image from Bag 6.

Fig. 3: Example images from the Udacity dataset.

(a) Sharp turns, intersections. (b) Narrow and long corridors.

Fig. 4: Indoor image examples acquired by the PMD.

3) Network III: UDMT: To keep the neurons specific to speed and steering separately, the last 2 layers were branched with each having 2 fully connected layers of 16 neurons and 1 neuron respectively; one for steering angle prediction and the other for speed prediction as in Fig. 2.

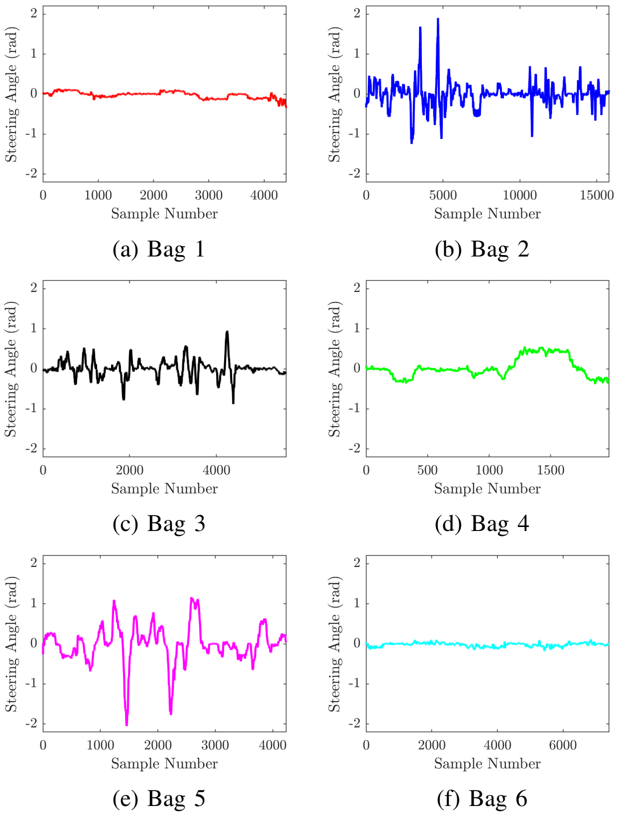
A. Network Parameters Learning

The *loss function* used is the mean squared error, which would result in a higher error if the predictions are further away from their actual values. Both steering and speed prediction were given equal weights in terms of their contribution to the final MSE error, unlike other works where special emphasis was entertained for the steering angle prediction by using a higher weight for its loss [16].

Different *optimisers* were considered during training; the classical stochastic gradient descent (SGD) procedure, Adadelta, ADAM and AMSGrad. In terms of convergence, AMSGrad [21] performed best. The convergence aspect of optimisers is constantly being revisited and it is worth noting new methods such as AdaMax [22], inspired by AMSGrad.

B. Data

Udacity [14] challenge 2 is a challege launched by Udacity to predict steering angles from camera images. This dataset is publicly available and widely referred in the autonomous driving community. Two sample images from this dataset are shown in Fig. 3. Even though the ultimate purpose of the research is to develop a framework for personal mobility, due to the similarity of the task it was decided to use



(a) Bag 1

(b) Bag 2

(c) Bag 3

(d) Bag 4

(e) Bag 5

(f) Bag 6

Fig. 5: Steering angles of bag files.

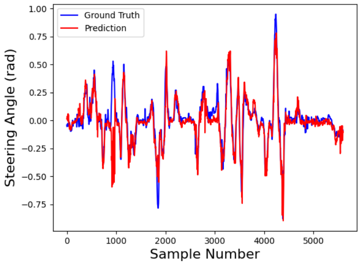


Fig. 6: Test on bag 3.

this dataset. Moreover, the trained network can then be used to infer the steering angle of the wheelchair collected in a completely different (indoor) environment to deduce its ability to generalize to unseen scenarios and vehicle dynamics. Two sample images from this dataset are shown in Fig. 4.

Since the accuracy of all the three models must be tested against one dataset for comparison, the general procedure of using the first five bags Udacity has provided (with a total of 33808 image frames) for training and using Bag 3 (with 5614 image frames) for testing cannot be done because Bag 3 only supplies images and steering. The proposed framework requires speed as well. Thus, after analysing the variation of steering angle in the six bags as in Fig. 5, 10000 - 15000 sample set of Bag 2 was set aside for testing and the other samples (excluding Bag 3) were used for training and validation. This resulted in a reduction in the dataset size. To overcome this, *data augmentation* was used, a technique widely used in the deep learning community to address the problem of having less amount of data and to make the trained network more robust and have generalization capabilities. Since storing the augmented data would result in high memory consumption, they were generated on the fly. Based on a random number generator, the images fed to the network were augmented to have different brightness and contrast levels and different noise levels.

TABLE II: Results on Udacity and PMD test sets

Network	RMSE for Udacity	RMSE for PMD
Image Only	0.1171 rad (Fig. 7a)	0.2228 rad (Fig. 8a)
UDM	0.0507 rad (Fig. 7b)	0.0725 rad (Fig. 8b)
UDMT Steering	0.0371 rad (Fig. 7c)	0.0555 rad (Fig. 8c)
UDMT Speed	0.1192 m/s (Fig. 7d)	0.1079 m/s (Fig. 8d)

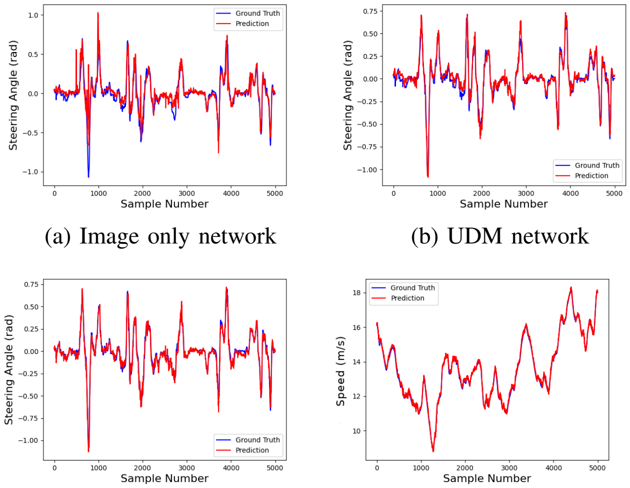
V. Results

The network in [8] was replicated on the same dataset reported by the authors (Udacity bag 3) to establish a baseline to validate the implementation (see Fig. 6). A root mean squared error (RMSE) of 0.1079 rad was achieved, while the authors reported 0.0613 rad. It is worth noting that beyond details not reported, the authors' aim is predicting steering at t based on images from $t - n + 1$ to t, whilst the present work strives to infer steering at $t+1$ considering the same sequence of input images. To test the accuracy of the proposed frameworks, the RMSE of the steering angle and speed at time $t + 1$ of the Udacity test set are considered. Results are summarised in Table II and graphed in Fig. 7.

In order to test how well the trained UDMT network can transfer to predict in novel environments and other mobility platforms, a new dataset was collected using the PMD shown in Fig 1. An Intel RealSense camera was used to collect the images. It was driven indoors around a typical office layout an altogether less dynamic environment consisting of narrow corridors, sharp turns and intersections (see Fig. 4) for around 15 minutes. The collected dataset consists of 21605 image frames and joystick command values. The models previously trained on Udacity data were further trained using 17000 images from the newly collected dataset. The speed commanded by the user along with the commanded steering was used as training inputs to the network instead of a car's linear speed supplied in the Udacity dataset. Results of the transfer learnt network are collected in Table II and Fig. 8.

VI. Discussion and Conclusion

A shared navigation framework to act based on inferred joint driver-machine intention has been presented in this work. It has been shown how adding user input and vehicle state to the image-only network improved the accuracy of steering angle prediction at time $t + 1$ by nearly 57%. Moreover, prediction of both steering and speed in the UDMT framework resulted in more accurate predictions. In terms of steering, speed can be thought of as privileged information which improves the accuracy of the steering prediction [23].



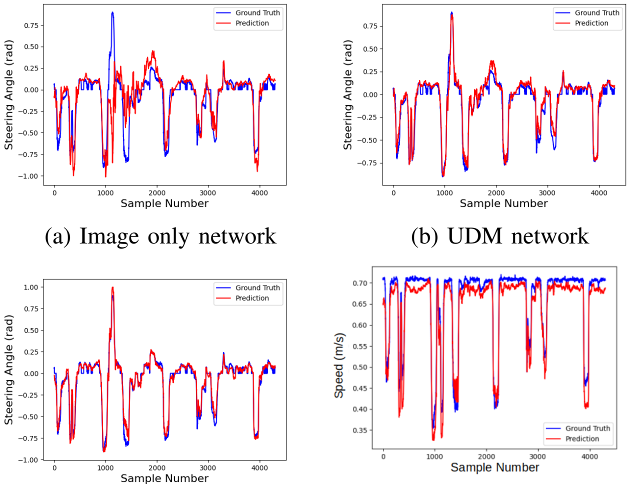
(a) Image only network

(b) UDM network

(c) UDMT steering

(d) UDMT speed

Fig. 7: Test results on Udacity dataset



(a) Image only network

(b) UDM network

(c) UDMT steering

(d) UDMT speed

Fig. 8: Test results on PMD dataset.

Moreover, the trained network was able to transfer knowledge to a new platform (PMD as opposed to car) and environment (indoor as opposed to outdoor) with only an additional small dataset of less than 12 minutes of training data. The accuracy of steering (and speed in the UDMT case) increased with the proposed networks, as was observed with the Udacity dataset. The computational time is proportional to the length of the input history, but since predictions are made at each time step, an online implementation gives a constant time per prediction.

For future work, additional user (accelerator/brake, gear) and vehicle states (pose of the vehicle) are being considered as inputs to the UMDT network. A new loss function to capture the smoothness of the joint control in addition to the error in prediction is also desirable. Uncertainty of the predicted joint control action will be estimated for reliability. Furthermore, the current framework gathers knowledge from a static database during training; methods to "learn to learn" such as meta learning will be further explored.

References

[1] M. A. Goodrich and A. C. Schultz, "Human-Robot Interaction: A Survey," *Foundations and Trends® in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[2] A. Bauer, D. Wollherr, and M. Buss, "Human-robot collaboration: A survey," *International Journal of Humanoid Robotics*, vol. 5, no. 1, pp. 47–66, 2008.

[3] J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction," *Journal of Human-Robot Interaction*, vol. 3, no. 2, p. 74, 2014.

[4] R. Dowling, J. Irwin, I. Faulks, and R. Howitt, "Use of personal mobility devices for first-and-last mile travel: The Macquarie-Ryde trial," *Proc of the 2015 Australasian Road Safety Conference*, 2015.

[5] M. E. Pollack, "Intelligent Technology for an Aging Population: The Use of AI to Assist Elders with Cognitive Impairment," *AI Magazine*, vol. 26, no. 2, pp. 9–9, 2005.

[6] M. Montemerlo, J. Pineau, N. Roy, S. Thrun, and V. Verma, "Experiences with a mobile robotic guide for the elderly," in *Proceedings of the National Conference on Artificial Intelligence*, 2002, pp. 587–592.

[7] D. P. Losey, C. G. McDonald, E. Battaglia, and M. K. O'Malley, "A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction," 2018.

[8] T. Wu, A. Luo, R. Huang, H. Cheng, and Y. Zhao, "End-to-End Driving Model for Steering Control of Autonomous Vehicles with Future Spatiotemporal Features," in *IEEE International Conference on Intelligent Robots and Systems*, 2019, pp. 950–955.

[9] C. Urdiales, A. Poncela, I. Sanchez-Tato, F. Galluppi, M. Olivetti, and F. Sandoval, "Efficiency based reactive shared control for collaborative human/robot navigation," in *IEEE International Conference on Intelligent Robots and Systems*, 2007, pp. 3586–3591.

[10] J. Leaman and H. M. La, "A Comprehensive Review of Smart Wheelchairs: Past, Present, and Future," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 486–489, 2017.

[11] S. P. Levine, D. A. Bell, L. A. Jaros, R. C. Simpson, Y. Koren, and J. Borenstein, "The NavChair Assistive Wheelchair Navigation System," *IEEE Transactions on Rehabilitation Engineering*, vol. 7, no. 4, pp. 443–451, 1999.

[12] C. Urdiales, J. M. Peula, M. Fdez-Carmona, C. Barrué, E. J. Pérez, I. Sánchez-Tato, J. C. Del Toro, F. Galluppi, U. Cortés, R. Annichiaricco, C. Caltagirone, and F. Sandoval, "A new multi-criteria optimization strategy for shared control in wheelchair assisted navigation," *Autonomous Robots*, vol. 30, no. 2, pp. 179–197, 2011.

[13] V. K. Narayanan, A. Spalanzani, and M. Babel, "A semi-autonomous framework for human-aware and user intention driven wheelchair mobility assistance," in *IEEE International Conference on Intelligent Robots and Systems*, 2016, pp. 4700–4707.

[14] Udacity, "The Udacity open source self-driving car project." [Online]. Available: https://github.com/udacity/self-driving-car

[15] A. Oussama and T. Mohamed, "A Literature Review of Steering Angle Prediction Algorithms for Self-driving Cars," in *Advances in Intelligent Systems and Computing*, vol. 1105 AISC, 2020, pp. 30–38.

[16] L. Chi and Y. Mu, "Deep Steering: Learning End-to-End Driving Model from Spatial and Temporal Visual Cues," *Proc of the Workshop on Visual Analysis in Smart and Connected Communities*, pp. 9–16, 2017.

[17] Z. Yang, Y. Zhang, J. Yu, J. Cai, and J. Luo, "End-to-end Multi-Modal Multi-Task Vehicle Control for Self-Driving Cars with Visual Perceptions," in *Proc - International Conference on Pattern Recognition*, 2018, pp. 2289–2294.

[18] F. Codevilla, M. Miiller, A. Lopez, V. Koltun, and A. Dosovitskiy, "End-to-End Driving Via Conditional Imitation Learning," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 4693–4700.

[19] S. Chowdhuri, T. Pankaj, and K. Zipser, "MultiNet: Multi-modal multi-task learning for autonomous driving," in *Proc - IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 1496–1504.

[20] V. Thakkar, S. Tewary, and C. Chakraborty, "Batch Normalization in Convolutional Neural Networks - A comparative study with CIFAR-10 data," in *Proc of 5th International Conference on Emerging Applications of Information Technology*, 2018.

[21] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *6th International Conference on Learning Representations - Conference Track Proc*, 2018.

[22] T. T. Phuong and L. T. Phong, "On the Convergence Proof of AMSGrad and a New Version," *IEEE Access*, vol. 7, pp. 61 706–61 716, 2019.

[23] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3530–3538.