Vision-based Steering Angle Prediction by the Fusion of Depth and Intensity Deep Features

Conference Paper · December 2018

CITATIONS
0

READS
83

1 author:

Vijay John
Toyota Technological Institute
48 PUBLICATIONS **340** CITATIONS

SEE PROFILE

Vision-based Steering Angle Prediction by the Fusion of Depth and Intensity Deep Features

ABSTRACT

In monocular camera-based end-to-end driving, the vehicle driving parameters such as steering angle, speed etc are directly estimated from the camera using deep learning. On the other hand, in traditional autonomous driving, these parameters are estimated using multiple modules such as sensing, behaviour generation, path planning and control. Owing to its ability to directly estimate the driving parameters, the end-to-end driving framework has received significant attention from the research community. In this paper, we present a novel stereo-based deep learning framework for end-to-driving, where the depth and appearance information generated using the stereo camera, are integrated to improve the steering angle prediction accuracy, especially for varying illumination conditions. Validation of the proposed algorithm is performed using multiple sequences of pre-defined driving routes with an expert driver. Each pre-defined driving route is acquired over multiple days with varying illumination conditions. Utilizing the acquired dataset, we show that the steering angle prediction accuracy of stereo-based end-to-end driving is better than monocular camera-based end-to-end driving.

KEYWORDS

Stereo vision, End-to-End driving

ACM Reference Format:

. 2018. Vision-based Steering Angle Prediction by the Fusion of Depth and Intensity Deep Features. In *Proceedings of ICVGIP (ICVGIP'18).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

2 INTRODUCTION

In recent years, rapid progress in the autonomous driving research community has been witnessed in order to improve the safety of the driver and the road users. The increase in safety is typically achieved by the intelligent processing of information received from multiple sensors, such as camera, GPS, LIDAR, etc, mounted on the vehicle. Among these different sensors, the inexpensive camera is widely used as it provides rich descriptive environmental information.

The vision sensors play an important role in the sensing module of the traditional autonomous driving framework. Traditional autonomous driving frameworks, widely used in literature, contain multiple modules for sensing, localization, path planning and control. However, this framework is susceptible to the limitations of the individual modules. For example, in case of GPS-based localization, the autonomous driving framework depends on the presence of strong GPS signals to localize the vehicle [23]. However, it is not always possible to receive strong GPS signals, especially in tunnel areas or high-rise building canyon. In such areas, the reliability of the autonomous driving framework decreases.

In recent years, researchers have used the deep learning framework to significantly advance the literature for autonomous driving. In the vision-based driving, termed as end-to-end driving, the vehicle driving parameters such as steering angle, acceleration and brake are directly predicted from the camera input using the deep learning framework. The end-to-end driving framework has the advantage of being used in scenarios where the traditional autonomous driving frameworks are reliable, like highway driving [11], as well as in scenarios where the traditional autonomous driving frameworks are less reliable. In both these scenarios, the end-to-end framework can be readily deployed, following training, as either a stand-alone driving framework or as a complementary framework for the traditional autonomous driving. In the latter case, the end-to-end driving framework could be used for first-mile or last-mile autonomous driving to complement the traditional driving framework. Owing to these potential advantages, this framework has received significant attention from the research community in the last few years [3, 11].

In end-to-end driving, the deep learning framework is trained to imitate the driver's behaviour captured using the vehicle CANBUS data and video information derived from the monocular camera. The monocular camera is used to generate appearance information of the road scene. The task of directly estimating the vehicle parameters from the monocular camera is not trivial. Some of the challenges include variations in the environment, road scenes and illuminations. These limitations are typically addressed in the literature by collecting large amounts of dataset with sufficient variation. An alternative solution to address these limitations is by performing sensor fusion of complementary information of the environment.

In this paper, we present a novel end-to-end autonomous driving framework for steering angle prediction. In this framework, an effective fusion of the appearance and depth information generated using a stereo camera is performed to overcome the illumination related limitation of monocular camera. The study is performed utilizing an end-to-end driving dataset acquired using an expert driver. The dataset, containing intensity, depth and steering angle information is acquired on multiple pre-defined routes acquired across varying illumination conditions (i.e morning, afternoon and evening). The results of the study show that the performance of the stereo-based end-to-end driving framework is better than the monocular camera-based end-to-end driving framework across multiple illumination conditions for similar pre-defined road scenes. To the best of our knowledge, the contribution to the literature are as follows:

- A novel appearance and depth-based end-to-end autonomous driving framework is proposed using the stereo camera.

ICVGIP'18, December 2018, Hyderabad, India

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

- Acquisition of dataset for stereo-based end-to-end driving acquired over multiple routes over different illumination conditions.

The remainder of the paper is structured as follows, the literature is reviewed in Section 3. In Section 4, the proposed deep learning framework is presented. The details of the dataset used for the study is presented in Section 5. The experimental results and detailed study is presented in Section 6. In Section 7, the paper is concluded with directions to future work.

3 LITERATURE REVIEW

Recently, with the rise of deep learning, end-to-end autonomous driving using monocular camera has emerged as an important research problem [4, 6, 7, 10, 11, 16, 22]. This is primarily due to the simplicity of the framework in directly predicting the vehicle driving parameters using the monocular camera input. Pomerleau et al. [16] proposed the ALVINN system, where an artificial neural network is used to predict the steering angles from gray-scale images. This approach was adopted and extended by Bojarski et al. [3] (NVIDIA) using the convolutional neural network. While reporting state-of-the-art results, the research problem is still unsolved due to challenges such as illumination variation, environment variation, non-linear steering trajectories etc. One possible approach to address these issues is to use a large number of labelled data containing significant illumination and environment variations, as observed in deep learning tasks such as image classification [12, 24]. To quickly obtain a large number of training data for end-to-end driving, researchers have proposed the use of synthetic data created with 3D rendering engines [18, 21]. Given, the generated synthetic dataset, the performance of the end-to-end framework is enhanced by utilizing learning [1] and control methods. [2]. However, the main drawback of the use of synthetic data is the limitation of their application to real-world scenes.

An alternative approach to address issues such as illumination variation is effective sensor fusion, which has been previously used for other deep learning-based ADAS tasks such as pedestrian detection [5, 20] and semantic segmentation [8, 9, 13]. In all these approaches, the depth information [5, 9, 13] or the thermal image information [20] are fused with the visible images to enhance the performance of ADAS tasks.

Consequently, in this paper, in the proposed end-to-end driving framework, the depth information is integrated with intensity information to enhance the steering angle prediction accuracy, even with limited real-world training data. Compared to literature, one of our main contributions is the use of depth and intensity information for end-to-end driving.

4 ALGORITHM

In this paper, a novel end-to-end driving framework is proposed to estimate the steering angle, x, directly from the stereo camera input, $\mathbf{z} = I, D$, where I represents the intensity image and D represents the depth image. The proposed end-to-end driving framework performs an inherent fusion of the intensity and depth information to estimate the steering angle. From the stereo rig, the intensity image I is obtained from the left camera, and the depth image D is obtained using the multipath Viterbi (MPV) algorithm [14].

The proposed framework is based on the convolution neural network (CNN) [12, 15, 17], which is considered the state-of-the-art for the environment perception tasks. The CNN is an end-to-end learning framework performing simultaneous feature extraction and regression. The features are extracted using multiple learnable layers of convolution filters, while the regression is performed using the dense network of fully connected neurons. The different weights and biases of the network are trained using the backpropagation algorithm.

In our CNN-based end-to-end algorithm, the feature extraction region extracts the features from the intensity and the depth image using two separate input branches. The features extracted from these separate branches are fused using a concatenation layer. The concatenated intensity and depth features are then given as an input to the regression region, which predicts the steering angle.

Each input branch of the feature extraction region contains multiple convolutional and pooling layers, with the initial and latter convolutional layers extracting the low-level and high-level features, respectively. The pooling layers perform the sub-sampling operation and reduce the dimension of the feature space. By representing the intensity and depth in separate branches, instead of representing them as a multi-channel single branch input, the convolutional layers extract more precise intensity and depth features. A detailed overview of the network parameters is presented in Fig 1. The network was trained with an Adam optimizer with a learning rate of 0.01, β_1 of 0.9, β_2 of 0.999 with no decay, batch size 20 and epochs 30. Henceforth, we refer to the proposed network as the "CNN-ID" network.

To validate the need for intensity and depth integration, we propose a variation of the CNN-based framework, the "CNN-I". In the "CNN-I", only the intensity-image is given as an input to the network. More specifically, in Fig 1, only the intensity branch is given as an input directly to the regression part.

5 DATASET

To facilitate the study for stereo-based end-to-end driving by the research community, we acquire a dataset with multiple pre-defined routes, R, as shown in Fig 2. For each pre-defined route r, multiple sequences, K, are acquired with varying illumination/environment conditions. The dataset with synchronized ZED stereo camera and vehicle CANBUS information is acquired using an experimental vehicle and an expert driver. The ZED camera is mounted outside the experimental vehicle, and is more susceptible to environmental conditions such as rain. The dataset is acquired with automatic gain and automatic white balance.

The dataset contains 4 pre-defined routes, represented as $R1, R2, R3, R4$. These routes were acquired over 2 days, represented as $D1, D2$. The acquisition was conducted thrice-a-day at the *Morning*, *Noon* and *Afternoon* time, respectively. Thus each pre-defined route has 6 sequences with varying illumination. The sequences for each route R are represented with a notation corresponding to the day and time of the acquisition (D_Time).

For example, for the route $R1$, $D1\_Morn$ in Fig 3 corresponds to a sequence acquired on $D1$ in the morning. The length of the sequence varies between 500-2500 frames depending on the route. Examples

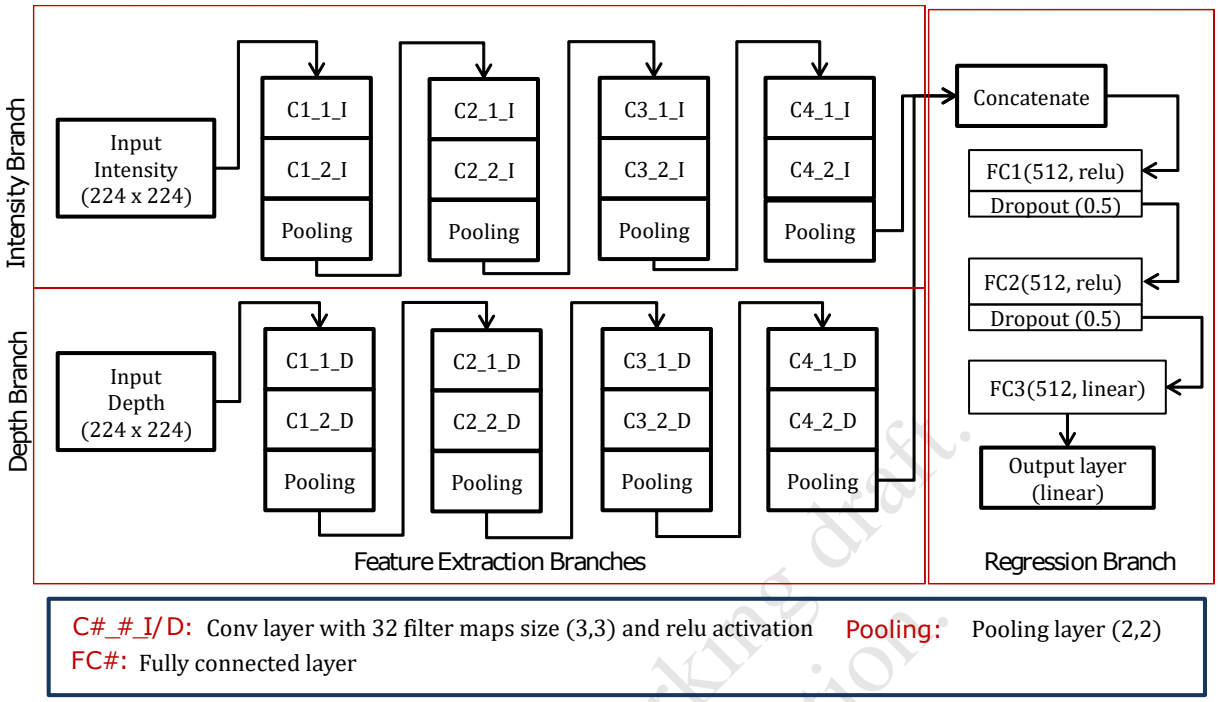


Figure 1: Detailed Architecture of CNN-ID network.

frames from the dataset illustrating the illumination variation are shown in Fig 2

Illumination Variation. The illumination statistics of the acquired dataset are calculated using the gray-scale intensity images of the stereo rig. For each route r and sequence k, with N intensity images, the **mean:** $\mu_r^k n$; **standard deviation:** $\sigma_r^k n$; **minimum value:** $\lambda_r^k n$ and **maximum value:** $\omega_r^k n$ are computed for each image n:

Subsequently, the image statistics for route r and sequence k are computed using the arithmetic mean as,

$$M_r^k = \frac{1}{N}\sum_{n=1}^{N} \mu_r^k n; \qquad \Sigma_r^k = \frac{1}{N}\sum_{n=1}^{N} \sigma_r^k n$$

$$\Lambda_r^k = \frac{1}{N}\sum_{n=1}^{N} \lambda_r^k n; \qquad \Omega_r^k = \frac{1}{N}\sum_{n=1}^{N} \omega_r^k n$$

The image statistics computed are illustrated in Figure 3, where the illumination variations are clearly observed for each route. These variations correspond to weather conditions observed during the data acquisition corresponding to “sunny" day, “cloudy" day and “rainy" day.

6 EXPERIMENTAL RESULT

In this paper, experiments are performed to validate the integration of depth and intensity information within the proposed framework for steering angle prediction. Our proposed algorithm is validated on the acquired dataset. The proposed algorithm is implemented on a Linux machine using Nvidia Geforce GT Titan X graphics card with Keras libraries. The experimental validation of the proposed framework is done by comparing with the baseline “CNN-I" algorithm and an end-to-end network based on the VGG-19 framework [19]. To facilitate fair comparison, the VGG-19 network’s inputs were changed from the colour image to the intensity image used in our experiments. The validation is performed by reporting the mean square error between the predicted steering angle and ground truth steering angle obtained from the vehicle CANBUS.

6.1 Comparative Analysis

In this experiment, we validate the proposed network and perform a comparative analysis with the baseline networks. For the comparative analysis, for each route, we train the different networks with five sequences and test the network on a single unseen sequence.

More specifically, we perform three trials for the comparative analysis of this experiment with three different unseen testing sequence, namely, D2-Afternoon, D1-Noon and D2-Morning. For each experimental trial, the networks were trained on five sequences and tested on the unseen sequence.

The weather for three testing sequences as shown in Fig 2, corresponds to *rainy*, *sunny* and *cloudy-drizzle*, respectively. The intensity statistics for the multiple training sequences are computed over all the images in the multiple training sequences.

Table 1: Mean square error for the comparative analysis on test sequence, (D2-Morning, Cloudy-Drizzle)

Alg.	R1	R2	R3	R4
CNN-I.	26.16	39.86	30.43	24.85
VGG-19 . [19].	30.08	43.63	28.96	30.76
CNN-ID	14.35	14.91	14.39	15.19

Road scene map

Road scene characteristics

Characteristics	R1	R2	R3	R4
Scene label	→	→	→	→
Road type	Private	Public	Public	Private
Shadows	High	Medium	Medium	High
Parked vehicles	High	Low	Low	High
Moving vehicles	Low	High	High	Low
Pedestrians	Medium	High	High	Low
Sky visible	Medium	High	High	Medium

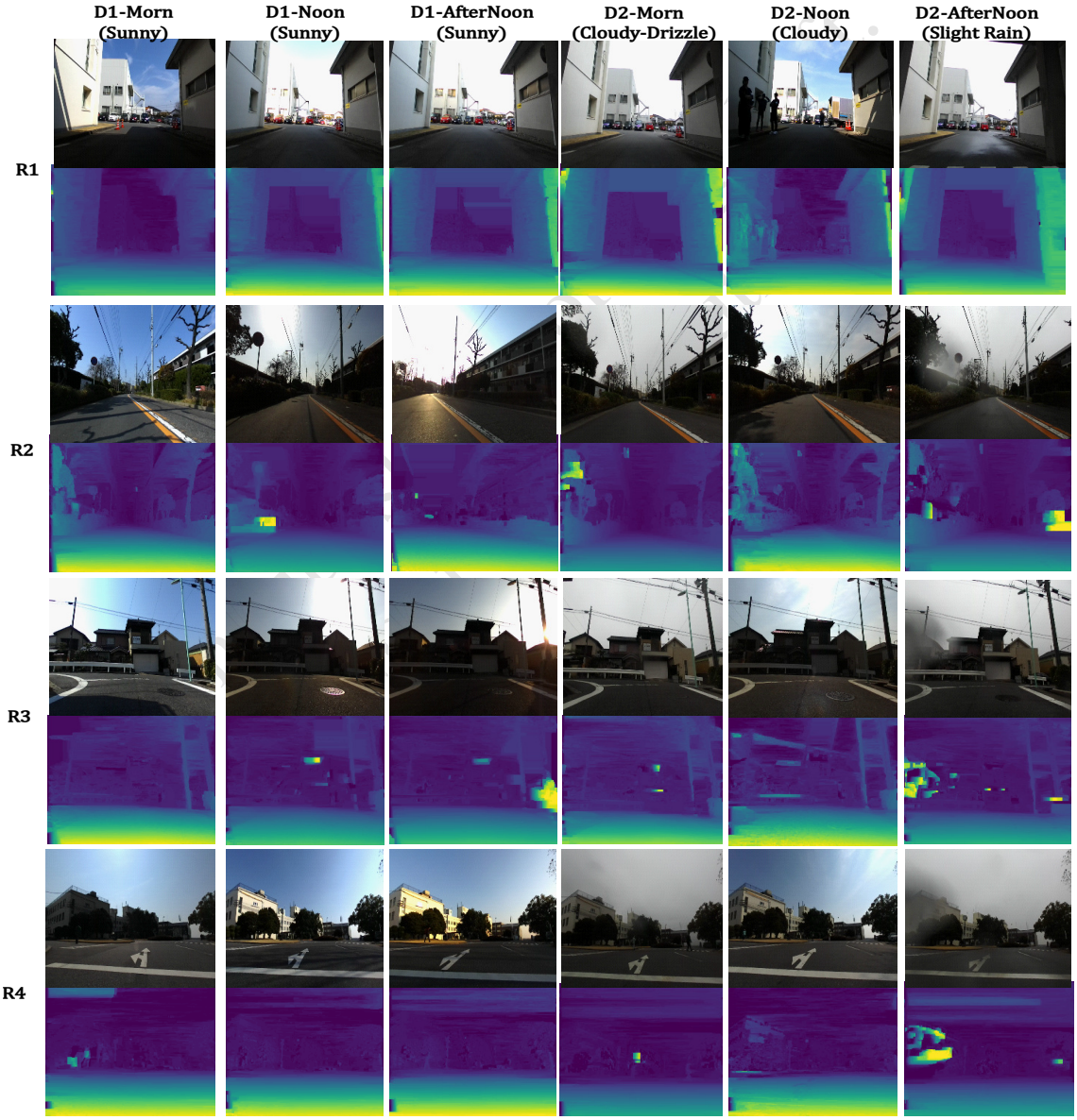


Figure 2: An Overview of the different routes in the dataset.

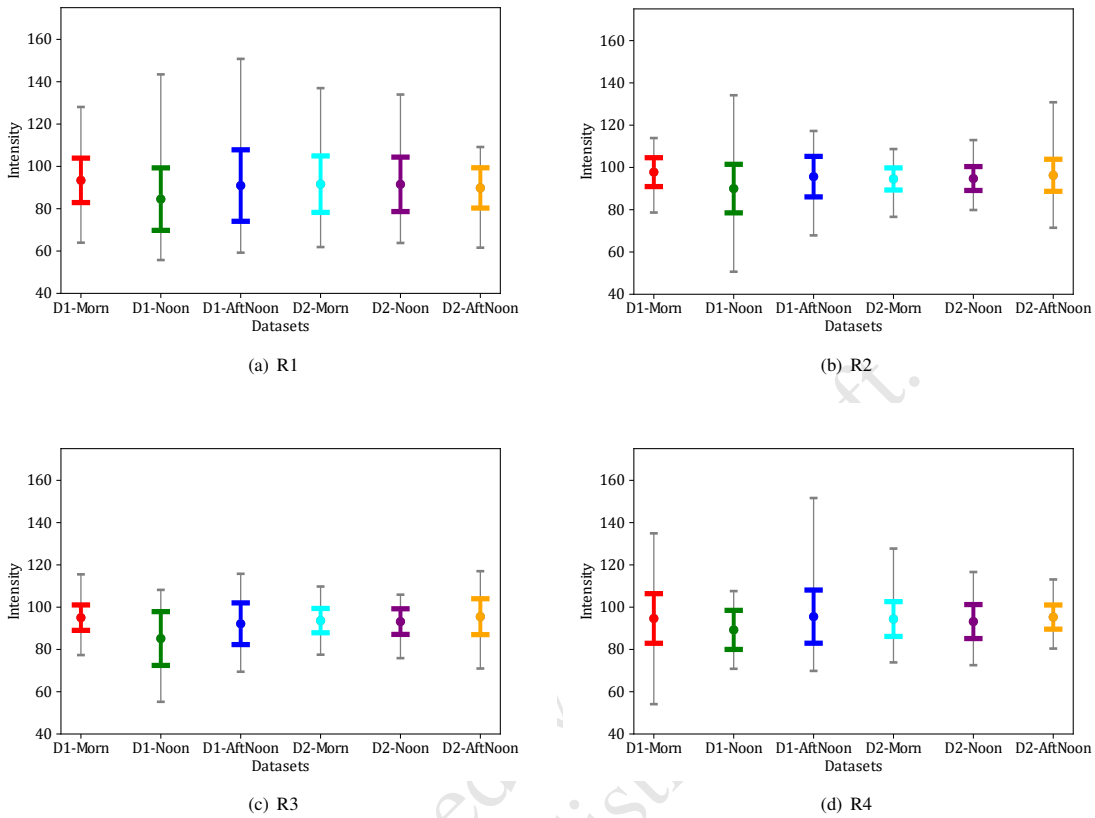


Figure 3: Intensity statistics for the different pre-defined routes. The colored error bars represent the mean and std.dev values, while the gray error bars denote the min and max values for each sequence.

Table 2: Mean square error for the comparative analysis on test sequence (D2-Afternoon, Rainy)"

Alg.	R1	R2	R3	R4
CNN-I.	46.41	40.83	22.75	34.96
VGG-19. [19].	51.73	69.71	41.13	51.73
CNN-ID	25.65	52.84	22.67	44.93

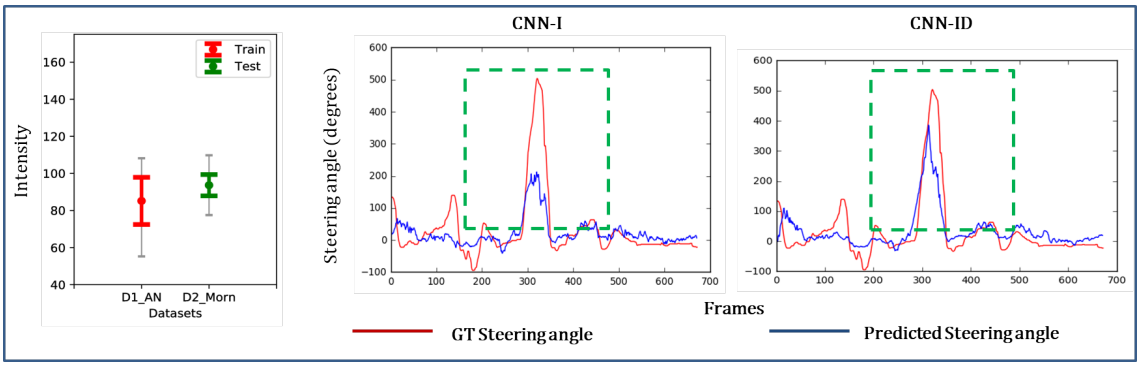
Table 3: Mean square error for the comparative analysis on test sequence (D1-Noon, Sunny)"

Alg.	R1	R2	R3	R4
CNN-I.	56.64	60.48	23.00	32.82
VGG-19. [19].	50.08	55.14	39.69	55.66
CNN-ID	21.18	14.32	10.42	19.03

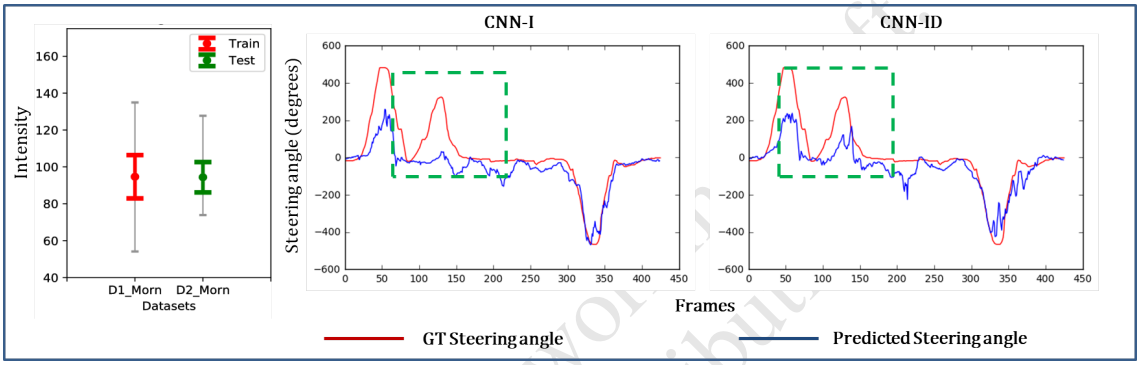
Discussion. In these experiments, we validated the network with varying illumination conditions corresponding to *sunny*, *cloudy-drizzle* and *rainy* weather. The results tabulated in Tables 1 and Table 3 show that for varying environmental conditions like *sunny* and *cloudy-drizzle*, the performance of the "CNN-ID" with the depth information is better than the performance of the intensity-alone baseline networks. Thus highlighting the advantages of integrating depth input features for robust end-to-end driving. However, in case of the *rainy* weather, the performance of all the three networks are inferior to the performances in other two test illumination condition. On closer inspection of the results, we also observe that the performance of the "CNN-ID" is similar to that of the intensity-alone networks (Table 2). This is primarily due to the noisy depth maps obtained in the rainy weather. The noisy depth maps are obtained due to the presence of rain drops on the camera lens, which are mounted outside the vehicle.

6.2 Parametric Analysis

In the comparative analysis experiment, as observed in machine learning literature, we trained the networks with multiple sequences and tested on an unseen sequence. In the parametric analysis experiments, we compare the performance of the "CNN-ID" and "CNN-I" networks when trained with a single training sequence, and tested on a single unseen sequence. We compare the results obtained with the single sequence trained network with the results obtained by the multiple sequence trained network. For each experiment, following training, the mean square error between the predicted steering angle and the ground truth steering angle is tabulated in Table 4-Table 7. Note that for the multiple sequence training, the networks were



(a) **R3-Sequence**. Ref Table 6



(b) **R4-Sequence**. Ref Table 7

Figure 4: Parametric analysis of the end-to-end network trained with single sequence. Green boxes denote the improved estimation accuracy for the "CNN-ID" for the dissimilar sequences.

trained with 5 different sequences, and tested on unseen sequence. Illustrations of the error graphs are provided in Fig 4

Table 4: Errors for the R1 route

Experimental Detail	CNN-I	CNN-ID
Multiple seq.training: Tested on $D2-AN$	46.41	25.65
Single seq.training: Trained on $D1-Morn$, Tested on $D2-AN$	77.97	67.97

Table 5: Errors for the R2 route

Experimental Detail	CNN-I	CNN-ID
Multiple seq.training: Tested on $D1-N$	60.48	14.32
Single seq.training: Trained on $D1-Morn$, Tested on $D1-Noon$	76.88	70.40

Table 6: Errors for the R3 route

Experimental Detail	CNN-I	CNN-ID
Multiple seq.training: Tested on $D2-M$	30.43	14.39
Single seq.training: Trained on $D1-AN$, Tested on $D2-M$	41.60	39.60

Table 7: Errors for the R4 route

Experimental Detail	CNN-I	CNN-ID
Multiple seq.training: Tested on $D2-M$	24.85	15.19
Single seq.training: Trained on $D1-Morn$, Tested on $D2-M$	76.9	71.58

6.3 Experimental Summary

Based on our study and the experimental results, we can observe the following:

- Illumination variation within the same pre-defined routes affects the performance of end-to-end driving.
- Integration of complementary features such as depth, improves the performance of end-to-end driving. However, the depth features are also susceptible to certain environmental conditions such as rain.
- As observed in the deep learning literature, a larger dataset with different illumination variations corresponding to multiple training sequences improves the performance of end-to-end driving.

7 CONCLUSION

In this paper, we present an end-to-end framework using stereo vision. In the proposed end-to-end network, the depth and appearance information generated using the stereo camera are fused. In the experiments, we show that variation in illumination affects the performance of intensity-based end-to-end driving, even for similar driving scenes. However, the addition of depth information from stereo reduces these errors. The effect of illumination variation and the advantages of the proposed network are demonstrated on an acquired dataset containing multiple trials of varying illumination on the same scenes. In the future work, we will investigate methods to further increase the robustness of end-to-end driving, and will deploy the method as a first-mile and last-mile driving framework to complement the traditional driving framework.

REFERENCES

[1] E. Perot A. El Sallab1, M. Abdou1 and S. Yogamani. 2016. Deep Reinforcement Learning framework for Autonomous Driving. In *NIPS Workshop*.

[2] M. Al-Qizwini, I. Barjasteh, H. Al-Qassab, and H. Radha. 2017. Deep Learning Algorithm for Autonomous Driving using GoogLeNet. In *IV*.

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. 2016. End to End Learning for Self-Driving Cars. *CoRR* abs/1604.07316 (2016).

[4] Lu Chi and Yadong Mu. 2017. Deep Steering: Learning End-to-End Driving Model from Spatial and Temporal Visual Cues. *CoRR* abs/1708.03798 (2017).

[5] O. Danut, A. Rogozan, F. Nashashibi, and A. Bensrhair. 2017. Fusion of stereo vision for pedestrian recognition using convolutional neural networks. In *ESANN*.

[6] S. Du, H. Guo, and A. Simpson. 2017. *Self-Driving Car Steering Angle Prediction Based on Image Recognition*. Technical Report CS231-626. Department of Computer Science, Stanford University.

[7] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2017. Going deeper: Autonomous steering with neural memory networks. In *CVPR*.

[8] Saurabh Gupta, Ross Girshick, Pablo ArbelÂt'aez, and Jitendra Malik1. 2014. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In *ECCV*.

[9] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. 2017. *FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture*.

[10] C. Hubschneider, A. Bauer, J. Doll, M. Weber, S. Klemm, F. Kuhnt, and M. Zollner. 2017. Integrating End-to-End Learned Steering into Probabilistic Autonomous Driving. In *ITSC*.

[11] V. John, S. Mita, H. Tehrani, and K. Ishimaru. 2017. Automated Driving by Monocular Camera Using Deep Mixture of Experts. In *IV*.

[12] A. Krizhevsky, I. Sutskever, and G. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

[13] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. 2016. LSTM-CF: Unifying Context Modeling and Fusion with LSTMs for RGB-D Scene Labeling. In *ECCV*.

[14] Qian Long, Qiwei Xie, Seiichi Mita, Hossein Tehrani, Kazuhisa Ishimaru, and Chunzhao Guo. 2014. Real-time Dense Disparity Estimation based on Multi-Path Viterbi for Intelligent Vehicle Applications. In *Proceedings of the British Machine Vision Conference*.

[15] A. Mahendran and A. Vedaldi. 2016. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision* (2016).

[16] Dean A. Pomerleau. 1996. Neural Network Vision for Robot Driving. In *The Handbook of Brain Theory and Neural Networks*.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.

[18] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. 2016. Playing for data: Ground truth from computer games. In *ECCV*.

[19] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[20] J Wagner, V. Fischer, M. Herman, and S. Behnke. 2016. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*.

[21] Bernhard Wymann, Christos Dimitrakakis, Andrew Sumner, Eric Espié, Christophe Guionneau, and Rémi Coulom. 2013. TORCS, The Open Racing Car Simulator, v1.3.5. `http://www.torcs.org`.

[22] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. 2017. End-to-end Learning of Driving Models from Large-scale Video Datasets. (2017).

[23] Yuquan Xu, Vijay John, Seiichi Mita, Hossein Tehrani Niknejad, Kazuhisa Ishimaru, and Sakiko Nishino. 2017. 3D point cloud map based vehicle localization using stereo camera. In *IVS*.

[24] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database.. In *NIPS*.