2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Vision and Dead Reckoning-based End-to-End Parking for Autonomous Vehicles

S. Rathour[1], V. John[1], M. K. Nithilan[1], S. Mita[1]

***Abstract*— In this paper, a combined vision and dead reckoning-based parking system for end-to-end driving are proposed. Standard autonomous parking frameworks contain multiple modules with each module having its own limitation. On the other hand, the proposed parking framework consists of a single end-to-end module, which reduces these inherent limitations. In the proposed deep learning-based parking system, a novel iterative two-stage learning framework is utilized to predict the steering angles and gear status using a front and back mounted monocular camera. In the first stage of the proposed framework, the encoder-decoder architecture is used to predict an initial estimate of the steering angle trajectory from multiple frames of the front or the back monocular camera. The camera used for steering estimated is selected using the gear status estimate. The gear status is predefined during initialization and estimated subsequently in the second stage of the proposed framework. In the second stage of the proposed framework, the initial estimate of the steering angle trajectory along with the vehicles heading angle, and absolute position is given as an input to the long short-term memory network to estimate the optimal steering angle and gear status. The proposed framework is validated on an acquired dataset. A comparative analysis of baseline algorithms and detailed parametric analysis are performed. The experimental results show that the proposed framework is better than the baseline end-to-end algorithms.**

I. INTRODUCTION

Significant achievements have been made in the field of autonomous driving and advanced driver assistance systems (ADASs). Most of the major automobile manufacturers have adopted intelligent parking assist system (IPAS), bringing parking assist and semi-autonomous parking as a mainstream automobile technology in the market. The earlier parking assists technology to have the ability to reverse park into an empty parallel parking space using a single manoeuvre(with $\pm 70cm$ front and back clearance). These systems also have the ability to detect the parking space from distances of $0.5 - 1.5m$. Present IPAS technology, an upgrade of the previous version, are capable of multi manoeuvre(with $\pm 40cm$ front and back clearance); reverse/forward parking into bay parking space. In spite of being commercially available, the parking assist system has inherent limitations and have not completely solved the parking problem. Moreover, they require human intervention at some level [2], [3] as well as structured parking lots. Hence, in order to develop the fully autonomous self-parking system this paper introduces combined vision and dead reckoning-based (DR) parking system. The main objective of the paper is to develop a learning-based framework to overcome the limitation of the present autonomous parking system (dependency on the structured parking lot, adjacent vehicle etc.). The main contribution of the paper is as follows:

- A novel two-stage deep learning based end-to-end parking system(Figure 2). capable of parking in structured as well as unstructured parking areas.
- Use of encoder-decoder architecture for end-to-end driving. In baseline end-to-end driving frameworks [Section IV], the features extracted from the image are given as input to the regression network. On the other hand, in the encoder-decoder architecture, salient features in the decoder output map are given as input to the regression network. This is shown to improve the estimation accuracy (Section V).
- Use of vehicle heading and distance travelled derived from DR to enhance the prediction accuracy of the proposed deep learning framework. (Sec III)

The remainder of the paper is divided as follows. Section II gives an overview of review work and limitation of the present parking technology. Section III delineates the proposed algorithm developed for end-to-end parking. Section IV explains about the dataset preparation for training and validation of the proposed deep learning based end-to-end parking algorithm. Section V gives a comparative study of the proposed learning framework with other baselines end-to-end learning framework as well as parametric variation of the proposed framework. Finally, in Section VI, the paper is concluded by listing the main contributions of the paper.

II. LITERATURE SURVEY

Most of the major automobile manufacturers (e.g. Toyota, BMW, Ford, Volkswagen, Mercedes-Benz etc.) roll-out their automobile equipped with semi-autonomous parking or IPAS system. The fundamental across the various semi-autonomous or IPAS technology remains similar. Firstly, the system begins with detection of suitable parking space. Next, the system formulates the best approach to get into the space informing the driver about the safe distance of obstacles around the vehicle. In order to perform the above-mentioned parking procedure semi-autonomous parking or IPAS system consists of multiple range based sensor and camera, mounted on the front and rear bumper of a vehicle to detect most of the variable involved. Hence, the parking system can be divided into multiple sub-modules such as environment perception, path generation, control and collision avoidance. Each of these modules is challenging research problems by themselves and require individual attention by the research

[1] S. Rathour, V. John, M. K. Nithilan and Seiichi Mita are with the Toyota Technological Institute, Japan {rathour,vijayjohn,nithilan,smita}@toyota-ti.ac.jp.

Fig. 1: A collection of images taken by front camera (top row) and back camera (middle row) for training end-to-end parking. (Last row) Image of the respective parking lot captured from hand held camera.

community. As per the sensors used semi-autonomous or parking assist systems can be divided into active sensor-based parking (ultrasonic or laser-based), vision-based systems or a combination of vision and active range sensors [4]. Ultrasonic or laser-based parking assist are most common [5], [6] sensor used for parking assist, however, they are prone to their own limitation [7]; for example, the ultrasound sensor-based systems are limited to short ranges and have difficulties in perceiving certain objects in the environment. In addition, laser-based range sensor is accurate, however, their high cost and short life limit their use. On the other hand, vision-based parking assistance systems require a structured environment like the presence of parking lane markers or the adjacent parked vehicle [8], [9], [10]. However, as observed in Fig. 1), often the parking lane markers are either missing or occluded and in some case, there is no adjacent parked vehicle. Additionally, vision systems are susceptible to illumination variation and environmental noise [11].

In this paper, we propose a vision and dead reckoning-based parking system using deep learning for autonomous vehicles to address the above-mentioned limitations of current semi-autonomous or parking assist systems. Recently there has been an increase in interest in the deep learning framework as it has provided state of the art results in the field of image recognition and segmentation with human-level accuracy [11], [12], [13] and autonomous driving [14], [15], [16], [17]. Compared to traditional autonomous driving frameworks, the deep learning-based frameworks consists of a single module or an end-to-end learning framework which directly predicts the steering angle from an image. Such systems eliminate the need for multiple modules such as environment perception, path planning, obstacle avoidance and control. End-to-end learning based algorithm requires minimal human effort being fully end-to-end trainable; taking image observation as input and steering control command as output. Hence, end-to-end learning is appealing, as it removes the need to explicitly model the different modules. However, existing literature on deep learning-based end-to-end driving is limited to highway and public road driving [14], [15], [16], [17]. In this work, we extend the end-to-end deep learning framework to autonomous parking. Additionally, we address the issues with traditional parking assist systems, especially the vision-based parking assistances which need a structured environment.

III. Algorithm

The main objective of this paper is to predict the sequence of expert steering angles and the gear status, given the time synchronized sequential image observations obtained using a front and back mounted fish-eye camera. Additionally, the distance travelled and vehicle heading (dead reckoning) are also used to predict the steering angles. The proposed network is a two-stage end-to-end learning framework consisting of an encoder-decoder stage and LSTM stage. Figure 2, shows the detailed architecture of the different stages of the proposed framework.

In the first stage, the encoder-decoder architecture based deep learning framework, used for semantic segmentation [18], is modified to predict the steering angle (Fig. 2) using images obtained from the front or back mounted fish-

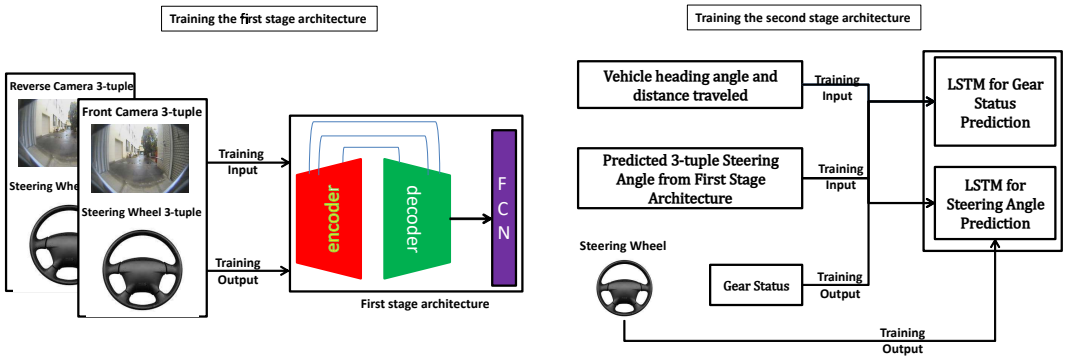

Fig. 2: An overview of the training step.

eye camera. The front camera images are used to predict the steering angles during the forward motion, and the back camera images are used to predict the steering angles during the reverse motion. The choice of camera is dependent on the gear status. For the initial frames, we assume the car to be moving forward and use the front cameras. For the subsequent framework, we estimate the gear status using the second stage of the proposed framework.

To predict the first stage steering angles, the deep segmentation network framework as proposed in [18] i.e. UNET was selected due to its simplicity and advantage over other baseline segmentation networks [12], [13]. As shown in Fig. 2, the first stage of the deep learning framework is formulated using the UNET and a refinement network. More specifically, to combine the two networks, the output of the U-Net decoder block is given as the input to the refinement network. The refinement network consists of 3 fully connected layers. First two fully-connected layers contain 512 units each followed by 256 in the last. The first two fully connected layer uses ReLu nonlinearity; whereas the last one uses linear activation function. Mean square error as loss model is used with Adam optimizer (1e-5) for learning the weights of the model.

In the second stage, the LSTM network is used to refine the first stage steering angle prediction and estimate the gear status. The input to the LSTM network is a sequence of stage one steering angle predictions along with the vehicle dead reckoning measurements. We next explain the training and testing steps of the algorithm in detail.

1) Training Step:

a) Encoder-Decoder Training: The encoder-decoder based fully connected network is trained using 3-tuple image observation i.e.(I_{i-1}, I_i, I_{i+1}) as training input and 3-tuple steering angle i.e. (s_{i-1}, s_i, s_{i+1}) as output. A typical parking manoeuvre consists of both forward driving and reverses driving. To enhance the accuracy of steering angle prediction, we utilize two separate cameras for image acquisition. During the forward manoeuvring, the 3-tuple image observations are acquired using the front mounted cameras. On the other hand, during the reverse manoeuvring, the 3-tuple image observations are acquired using the back-mounted cameras. Note that the 3-tuple image observations are always synchronized with the 3-tuple steering angles obtained from the vehicle CANBUS. The images acquired from both the cameras are used to train the first stage network.

Given the training data, the image-to-steering mapping function in the first stage network is trained in a supervised manner. More specifically, the parameter θ of the regression function approximated by proposed encoder-decoder based fully connected framework is optimized.

b) LSTM Training: In the second stage of the network, the dead reckoning (DR) measurements are used to enhance the performance of proposed deep learning framework by refining the first stage steering angle estimates. More specifically, the first stage steering angle estimates along with DR measurements are given as an input to the LSTM network which generates an optimal estimate of the steering angle. Additionally, the LSTM network also predicts the gear status, which is used during the testing phase for the first stage camera selection.

The DR measurements correspond to the distance travelled (i.e. d) by the vehicle; derived using time-synchronized can bus data comprised of four-wheel rotational speed, vehicle speed, gear status g and yaw rate $\dot{\psi}$. The proposed LSTM network takes predicted $\hat{s}$, from the first stage deep network along with the distance travelled d and absolute heading angle ψ. The LSTM network consists of two outputs, first is LSTM based classifier to predict the $\hat{g}$. As $\hat{g}$, is composed of two state only classifier based LSTM model is better suited for this task. Secondly, LSTM model predicts $\hat{s_f}$ with linear activation unit and mean square error loss model.

2) Testing Phase: The parameter obtained after training the two-stage proposed deep network framework is used to map the time synchronized I,d and ψ to $\hat{s_f}$ (final predicted control steering angle) and $\hat{g}$ (predicted gear status). The overview of the testing is shown in Fig. 3.

During initialization, the proposed algorithm is initialized with the assumption that the vehicle is moving forward with gear status $g = 1$. Consequently, during initialization, the first stage network uses the front camera to estimate $\hat{s}$.

The initial estimate of the steering angle $\hat{s}$ along with the vehicle DR measurements are given as an input to the second stage LSTM network. The LSTM estimates the optimal

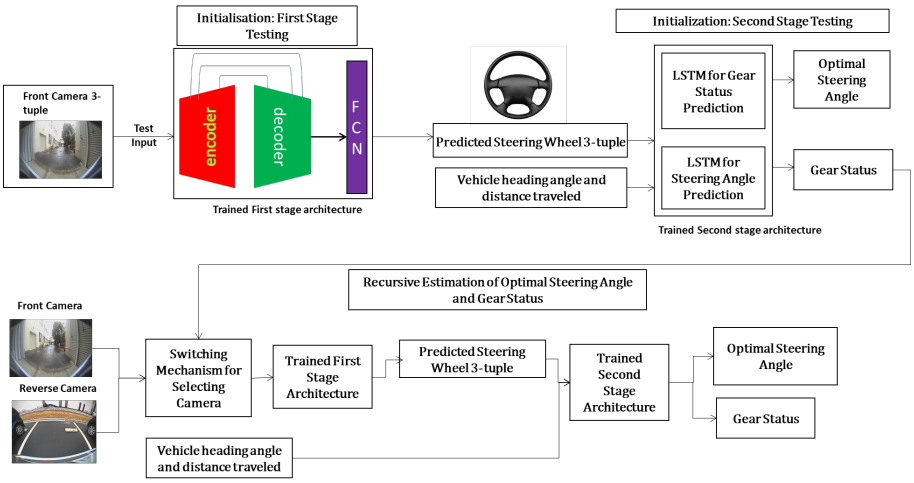

Fig. 3: An overview of the testing step.

steering angle $\hat{s_f}$ along with the gear status $\hat{g}$. The gear status is then recursively used by the first stage network to select the back or the front camera for the input. $g = 1$ implies front camera and $g = 0$ implies back camera.

IV. Experimental Results

To begin with, firstly, we introduce the dataset used for training and testing the proposed deep learning framework. Finally, we delineate the experimental setup and result from discussion. A comparative analysis is performed with baseline models. Additionally, we also perform a parameter analysis with variations of the two-stage network. All the training and validation was performed on the system with the following specification:- 64-bit Intel Core i7-6850K CPU @ 3.60GHz×12, GeForce GTX 1080, RAM 64 GB using keras with tensor flow backend.

A. Experimental Dataset

In order to learn end-to-end visual perception based deep learning framework for autonomous reverse parking, expert driver demonstrated parking was used to prepare the dataset for training and testing. The expert driver was asked to perform numerous parking to create the various dataset. Five datasets (Fig. 1) i.e. D1, D2, D3, D4 and D5 consisting of two sequences each was prepared for training and testing respectively. Each dataset consists of time synchronized front and back camera image observation I,ψ,s and d. Each dataset was partitioned into separate training and testing sequences. Figure 1 shows the image captured by front and back camera for all the five datasets. Dataset D1, D2 and D3 were prepared using unstructured parking lot i.e. the parking lot with degraded white line and missing white line(Figure 1, last row). Hence D1-D3 represents unstructured parking lot and D4-D5 represents structured parking lot dataset (well defined white lines (Figure 1, last row)).

In order to validate the performance of the proposed algorithm, the proposed model was trained on the training sequence of each dataset (i.e. D1....D5) and finally the optimized parameter θ obtained after training was used to test on the testing sequence of the corresponding dataset. For example, the optimized parameter θ derived by training the proposed model on training sequence of D1 was tested on the testing sequence of D1.

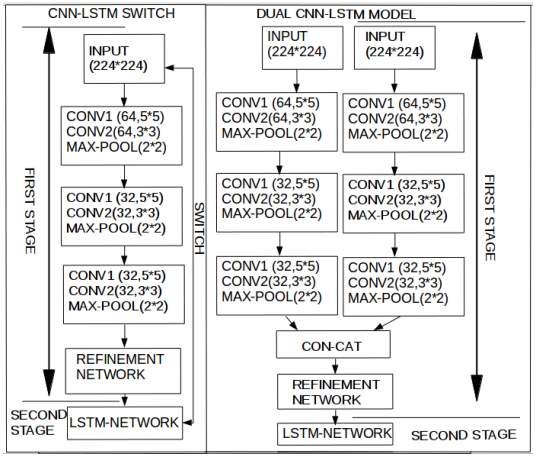

Fig. 4: CNN architecture of the variant deep learning based frameworks used for the parameter analysis.

B. Comparative Analysis

Four baseline deep learning based end-to-end networks were selected for the comparison with the proposed model. The CNN architecture used by [11] (i.e. VGG16) to map the image pixel to steering angle represents the first baseline. The residual based CNN architecture (RESNET50) as proposed in [20] represents the second baseline. For the third and fourth baseline, we extracted feature maps extracted from the final convolutional layer of the VGG-16 and RESNET50 to train an extra trees-based regressor (i.e.VGG16 + ET & RESNET50 + ET) [21].

To demonstrate the benefits of utilizing two cameras in the proposed framework, the baseline models were trained

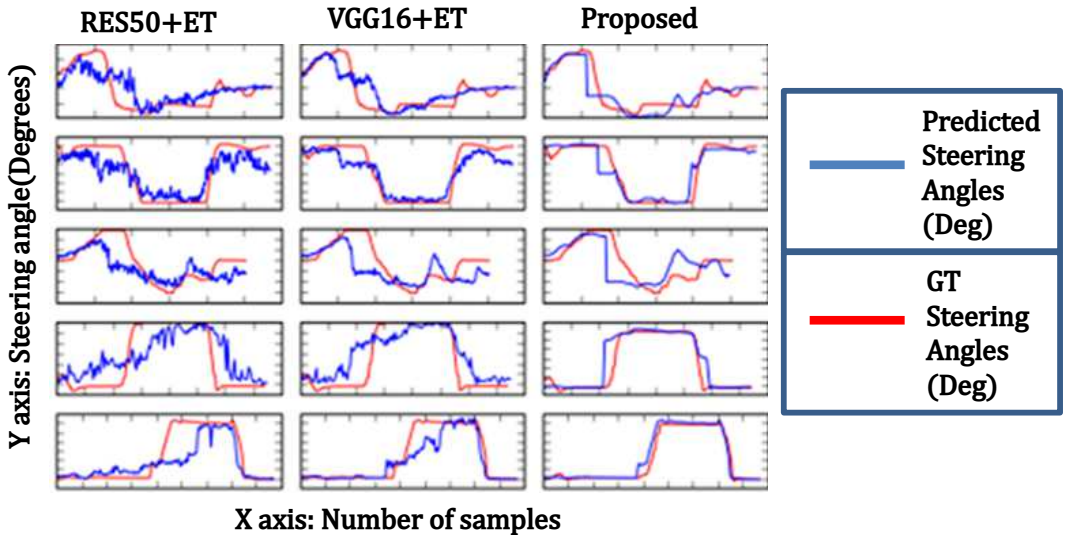

Fig. 5: Trajectories of predicted $\hat{s_f}$ (blue line) along with the ground truth trajectory s (red line) for each dataset (D1-D5), row-wise, using RES50+ET, VGG16+ET and Proposed.

TABLE I: Mean Euclidean distance (degrees) between the predicted and ground truth steering angles for the different datasets.

Data.	RES50	RES50+ET	VGG16	VGG16+ET	**Prop.**
D1	93.74	23.44	157.35	17.08	9.22
D2	64.02	56.00	30.13	41.0	7.89
D3	93.74	105.66	212.78	131.11	11.93
D4	64.47	43.75	157	65.35	38.61
D5	166.17	56.29	29.37	41.46	17.70

TABLE II: Mean Euclidean distance (degrees) between the predicted and ground truth steering angles for the different datasets

Data.	**Prop.**	CNN-LSTM-Switch.	Dual CNN-LSTM
D1	9.22	29.49	22.28
D2	7.89	28.28	31.69
D3	11.93	31.98	155.86
D4	38.61	54.54	133.25
D5	17.70	143.30	92.13

using the front camera "alone". Moreover, this is also done to imitate the automated driving networks used in the prior work [15], [21]. The performance of the four baseline models was studied on all the five datasets for comparison. The different models are quantitatively compared by measuring the mean Euclidean distance between the predicted steering angles and the ground truth steering angles.

The results obtained in Figure 5 and Table 1 show that the proposed network is better than the baseline algorithms across different datasets. Compared to the baseline, the improved performance can be attributed to the following:

- Two-stage architecture where the initial estimate of the steering angle is refined by the LSTM.
- Utilizing the encoder-decoder architecture to obtain an initial estimate of the steering angle, where the U-Net's output map is given as input to the fully connected network.
- Switching between two cameras for prediction.

C. Parameter Analysis

After the comparison of the proposed model with the respective baseline models, the proposed model architecture at the first stage was varied to study the parametric dependency. Two variations of the proposed framework were used in the parametric analysis (Figure 4). In the first variation, we eliminate the camera switching mechanism and replace the U-Net with two CNN branches for the front and back camera images. The features extracted by the CNN branches are concatenated and given to the refinement network in the first stage. The initial steering angle is refined by the LSTM in the second stage. This model is called as *Dual CNN-LSTM model*.

In the second variation, we replace the U-Net in the proposed framework with CNN to extract the image features, while retaining the switching mechanism. This model is termed as the *CNN-LSTM switching model*. The CNN architecture for the two variant models is shown in Figure 4.

The results obtained in the parametric analysis are shown in Fig 6 and Table 2, show that the proposed model is better than the parametric variations. The parametric analysis demonstrates the advantages of using the U-Net, the switching mechanism and the LSTM framework. The advantages of the U-Net is evident in the comparison with the CNN-LSTM-Switching model, as both these models are similar except the U-Net in the proposed model and the CNN in the variant model.

The advantages of the U-Net and the switching mechanism are evident in the comparison with Dual CNN-LSTM model, as the switching mechanism is used to select the front "or" back in the proposed model. On the other hand, both the camera images are used by the variant model.

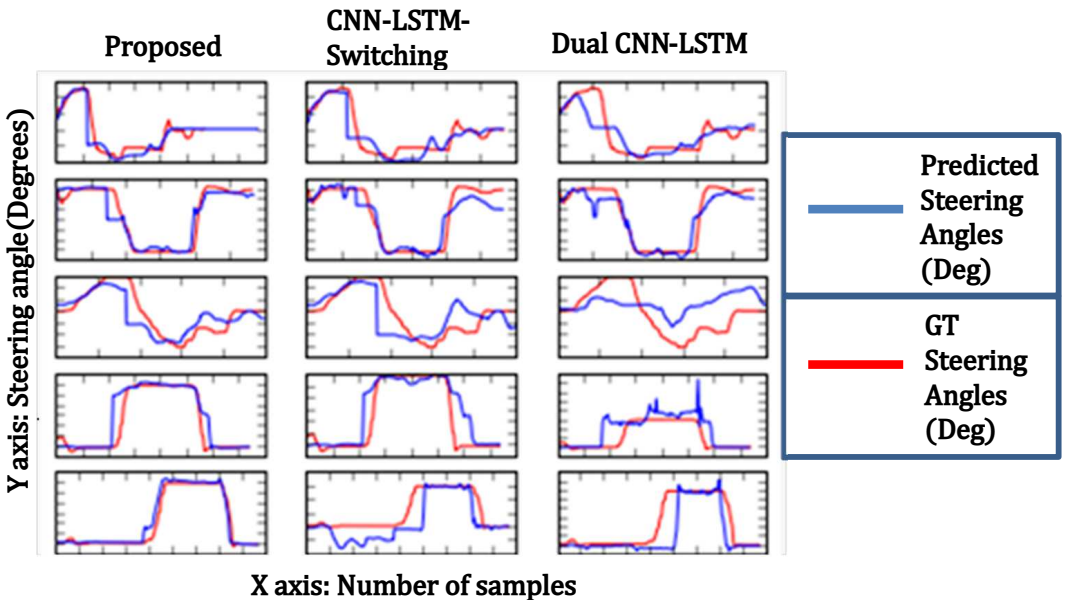

Fig. 6: Trajectories of predicted $\hat{s_f}$ along with the ground truth trajectory s for each dataset (D1-D5), row-wise, using proposed model, separate and combined model.

V. Conclusion

In this paper, a combined vision and DR based novel two-stage encoder-decoder architecture is proposed to predict the steering angles and gear status using front or back mounted monocular camera. The proposed model shows better performance compared to the baseline end-to-end models with limited training dataset. Few parameter variations were also performed at the first stage of the proposed network to study the effectiveness of the proposed model. The proposed model was found effective in predicting steering trajectory and gear status making the system fully autonomous and capable of multi man-oeuvre. The proposed deep learning based end-to-end parking can be used even for an unstructured parking lot with fully visible, partially visible/occluded parking white line or even in case if the parking lot has no parking line.

References

[1] W. Wang, Y. Song, J. Zhang, and H. Deng, "Automatic parking of vehicles: A review of literature," International Journal of Automotive Technology, vol. 15, no. 6, pp. 967-978, 2014.

[2] The Hybrid That Started it All, Mar. 2014. [Online]. Available: http://www.toyota.com/prius/

[3] BMW 7 Series. Park Assist, Mar. 2013. [Online]. Available: http://www.bmw.com/

[4] X. Du and K. K. Tan," Autonomous reverse parking system based on robust path generation and improved sliding mode control," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 3, pp. 1225-1237, Jun. 2014.

[5] J. Pohl, M. Sethsson, P. Degerman, and J. Larsson, "A semi-automated parallel parking system for passenger cars," Proc. Inst. Mech. Eng. D, J. Autom. Eng., vol. 220, no. 1, pp. 53-65, Jan. 2006.

[6] H. G. Jung, Y. H. Cho, P. J. Yoon, and J. Kim, "Scanning laser radar based target position designation for parking aid system," IEEE Trans. Intell. Transp. Syst., vol. 9, no. 3, pp. 406-424, Sep. 2008.

[7] P. Degerman J. Pohl M. Sethson "Ultrasonic sensor modeling for automatic parallel parking systems in passenger cars," SAE 2007 World Congress & Exhibition, Detroit, MI, U.S.A., 16th19th April, 2007 .

[8] K. Fintzel R. Bendahan C. Vestri S. Bougnoux T. Kakinami "3D parking assistant system," Proc. IEEE Intell. Veh. Symp.¡/em¿ pp. 881-886 2004.

[9] N. Kaempchen U. Franke R. Ott "Stereo vision based pose estimation of parking lots using 3d vehicle models" ¡em¿Proc. IEEE Intell. Veh. Symp.¡/em¿ vol. 2 pp. 459-464 2002.

[10] C. Wang,, Hengrun Zhang, Ming Yang, Xudong Wang, Lei Ye and Chunzhao Guo. "Automatic parking based on a bird's eye view vision system," Advances in Mobility Theories, Methodologies, and Applications, vol. 2014 pp. 847406-1-847406-13 2014.

[11] Alex Krizhevsky, IlyaSutskever, and Geoffrey E Hinton,"Imagenet classification with deep convolutional neural networks,"in Advances in neural information processing systems (NIPS), 2012, pp. 1097-1105.

[12] M. Thoma, "A survey of semantic segmentation," CoRR, vol. abs/1602.06541, 2016.

[13] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. G. Rodriguez, "A review on deep learning techniques applied to semantic segmentation," CoRR, vol. abs/1704.06857, 2017.

[14] C. Chen, A. Seff, A. Kornhauser, and J. Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In Proceedings of the IEEE International Conference on Computer Vision, pages 2722-2730, 2015.

[15] 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.

[16] H. Xu, Y. Gao, F. Yu, and T. Darrell. End-to-end learning of driving models from large-scale video datasets. arXiv preprint arXiv:1612.01079, 2016.

[17] Lu Chi and Yadong Mu. "Deep Steering: Learning End-to-End Driving Model from Spatial and Temporal Visual Cues". In: arXiv preprint arXiv:1708.03798 (2017).

[18] Y. LeCun, U. Muller, J. Ben, E. Cosatto, and B. Flepp. Offroad obstacle avoidance through end-to-end learning. In NIPS, pages 739-746, 2005.

[19] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In MIC- CAI, pages 234-241. Springer, 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages,770778, 2016.

[21] Vijay John, Seiichi Mita, Hossein Tehrani Niknejad, Kazuhisa Ishimaru,"Automated driving by monocular camera using deep mixture of experts,IV 2017, 10.1109/IVS.2017.7995709.