EU-TP0398

Safe Vehicle Trajectory Prediction Using Deep Neural Networks and Camera Images

Dzmitry Tsishkou[1*], Rémy Bendahan[1]

1. IMRA EUROPE, France

220 rue Albert Caquot, 06904 Sophia-Antipolis, 33-493-957-391, tsishkou@imra-europe.com

Abstract

Vehicle trajectory control devices (ESP) allow drastic reduction of driving accidents. The next step in driving safety is to predict which trajectory a vehicle should follow in order to further reduce accidents. None of the existing methods have proved to be safe, smooth and fast enough. To achieve high level of trajectory prediction and additionally cope with human failures (when driver is stressed or distracted), we propose an intelligent driver assistance function that is trained to copy human decisions in all driving scenarios. Human driver could be alerted or corrected in case where his driving control decision disagrees with the prediction of the intelligent assistant. Camera images are the richest source of data available among today sensors that could be used to interpret the context of driving scenario and recognize surrounding objects within it. In this paper, we propose the first safe vehicle trajectory prediction by a deep neural network system that is trained to estimate safely both longitudinal and lateral position of host vehicle. We demonstrate that our system could predict with 95% accuracy how acceleration and wheel angle should be changed within next 1 sec by analysing only image data acquired from past two seconds. We also present how such system could be extended for prediction of future vehicle trajectory up to 7 seconds ahead with 5% error in positioning distance.

Keywords:

Deep learning, trajectory prediction, driver assistance.

Introduction

Improving driving safety is one of the most important problems in automotive world [1]. Highly automated component of vehicle’s lateral and longitudinal control function could be another future step in preventing accidents on roads [2]. Today’s approaches consist in combining high accuracy 3D NAVI map, vehicle dynamics data and bunch of sensors to detect obstacles and predict their positions. The accuracy of such approaches depends on the accuracy of their input, which is unstable due to the following:

- 3D maps are subject to road architecture change e.g. brand new roads, stopped obstacles, roadworks.
- Obstacle detection depends on the sensors’ ability e.g. LIDAR can detect

obstacles but hardly recognize them to predict their positions; camera-based object recognition is limited in poorly contrasted situations.

Therefore tolerance of today's algorithms of trajectory prediction that combine the above mentioned input with human-engineered decision making can barely guaranty driving safety without driver intervention. It requires high attention from the driver in case of rare driving scenarios or situations due to limited ability of humans to model all possible situations with explicit set of rules or algorithms. The lack of attention is one of the major causes of accidents. Thus, an intelligent system should reduce the requirements in driver' attention. It should assist the driver to focus attention or avoid unsafe vehicle control decision to reduce the number of accidents and increase road safety.

This work is focused on end-to-end paradigm, which doesn't require obstacle detection function or own localization in accurate 3D maps but uses camera image data as input (starting end) and produces vehicle control decisions (finishing end). By combining context and obstacle recognition such system may not require high quality image data; therefore it could be more robust in hard weather conditions compared to traditional object recognition based systems. One of the best candidates to realize such end-to-end system is to use deep neural networks [3] inspired by human brain data processing organization, which recently outperformed all other existing state-of-the-art approaches by recognizing 1000 different classes of objects. Their unprecedented capacity of interpreting both context and main objects within images leads to superior generalization and decision making capacity among machine learning tools. By feeding this network with enough of data one could expect them to learn how to predict own-vehicle trajectories by copying and aggregating experience of real human drivers, without explicitly learning concepts of obstacles or path planning. The proposed approach is based on the fact that human drivers are trained to decide own-vehicle trajectory, based on analysis of surrounding objects for everyday driving scenarios that is safe and fast. Past experience also allows drivers to safely tackle new situations. In next sections, we will demonstrate how one could train convolutional neural network (CNN) to predict the required changes of wheel angle and acceleration for safe trajectory control. In our system, we use only VGA images as input data acquired with forward looking 130degrees FOV camera at rate of 30fps for wheel angle and acceleration prediction in the 1 to 7 second time interval.

Previous Work on Vehicle Trajectory Prediction for Control

Systems proposed in [4], [5] are trying to provide driving support or own vehicle path prediction. They are matching current vehicle position (from GPS data) with Navigation map in order to understand the radius of curvature of road ahead. They are comparing the current road structure and own vehicle driving speed. If the driving speed is too high for the current road, then guidance (path change) or speed control is realized. These systems may improve driving safety. However, they are not optimum when navigation maps are not accurate which is often the case e.g. when a vehicle is stopped on the side of the road. In such case, these systems may require/suggest to slow down which is inappropriate e.g. when overtaking the stopped vehicle.

[6-9] are providing countermeasures to the work above by adding an obstacle detection module. In this way, they can add obstacles to the navigation map and take them into account by the path planning algorithm. They rely on obstacle sensing which known to be not accurate enough due to actual sensors performance. Unreliable obstacle detection will lead to wrong guidance or control of own path (same issue as for [4], [5]). Overall the prediction of object motion is also unreliable due to the impossibility to predict in a time horizon greater than 1sec, the future position of objects especially of pedestrians that can change quickly their walking behaviour. Such lack of accuracy in object prediction may impact own path planning which can become unsafe when a pedestrian suddenly decides to cross a road.

The authors of [10] proposed a system that uses motor signals to provide unsupervised regularization in convolutional neural networks to learn visual representations from video. Specifically, the authors enforce that the learned features exhibit equivariance i.e. they respond predictably to transformations associated with distinct ego-motions. This system has demonstrated to achieve high accuracy in matching visually similar camera images captured at different times.

Comparison between end-to-end (contribution of this paper) and semantic abstraction approaches (current state-of-the-art) for vehicle control are discussed in [11]. In case of the end-to-end system we expect that given enough training data the deep neural networks could generalize well to copy experience of multiple drivers in everyday scenarios. Such deep neural networks would only receive visual input from camera images as a real driver and decide how vehicle’s trajectory within next few seconds. This approach is different from the semantic abstraction, where the same problem is decomposed by human engineers into many sub-problems, and the vehicle is controlled by an explicitly written algorithm. The potential advantage of the end-to-end approach is the ability to make fast and safe driving decisions equally well in both frequent and rare driving scenarios, while in case of human engineered algorithms we could expect lower performance in case of previously unseen or rare situations, due to absence of explicit rules that should be used in those cases.

Deep Learning based Single Control Point Prediction

The problem of lateral and longitudinal position prediction of a single control point (future trajectory composed of only one point) was formulated in the following way: by using sequence of images from T-2 sec. till T sec. (now) as input, machine learning (convolutional neural network - CNN in case of this paper) should predict changes of angle and acceleration required to be adjusted for safe trajectory control at time T+1sec as illustrated in Fig. 1.

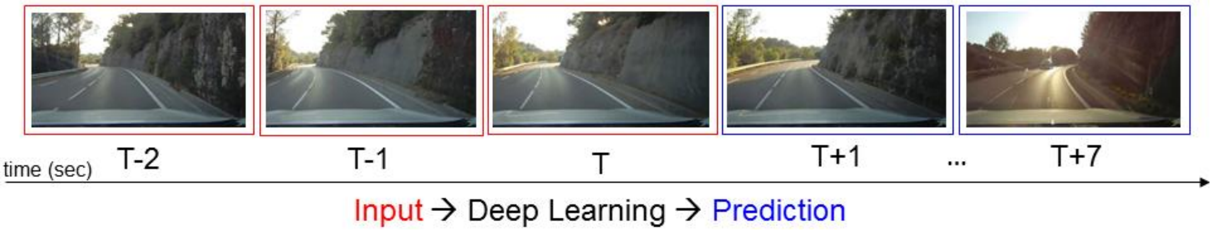


Figure 1 - Single control prediction problem

We decided to use a database of 10 hours of driving that includes images and GPS data. The data were recorded while driven by non-professional drivers in conditions without accidents. Since we collected data from multiple drivers, our system would learn trajectory of average driver, which is robust to individual driving errors coming from single data source. Dataset was acquired around Sophia Antipolis, France and is a representative of an everyday driving scenario from home to work and back. Such dataset can easily be acquired and automatically labeled by converting logged GPS data into vehicle dynamics data (speed and wheel angles). Our approach consists in using a CNN to predict acceleration and wheel angles at T+N sec using images as only input at real-time driving stage (Fig.2).

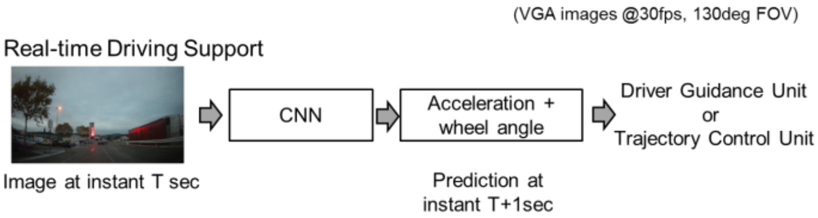

(VGA images @30fps, 130deg FOV)
Real-time Driving Support
CNN
Acceleration + wheel angle
Driver Guidance Unit or Trajectory Control Unit
Image at instant T sec
Prediction at instant T+1sec

Figure 2 – Real-time driving support based on CNN

At offline training stage (Fig.3), we use acceleration and wheel angle labels obtained from GPS data source synchronized with image data in order to generate the neural networks (CNN) that will be used during the real time driving support (Fig.2).

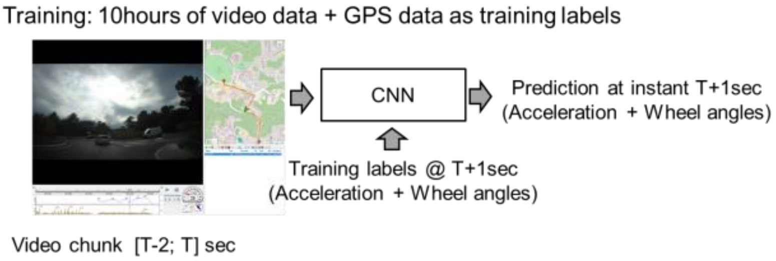

Training: 10hours of video data + GPS data as training labels
CNN
Prediction at instant T+1sec (Acceleration + Wheel angles)
Training labels @ T+1sec (Acceleration + Wheel angles)
Video chunk [T-2; T] sec

Figure 3 – CNN training for real-time driving support

One inherent advantage of such data as input to the deep neural network is the very simple way to accumulate large amount of data (just by setting-up a camera with data logger on the car and record data). In our approach, the labels used for the training are the GPS + Inertial data which are automatically acquired and synchronized with images. This makes data collection and processing step described in the next section fully automated.

Data Collection and Processing for Single Control Point Prediction

Data logger stores lateral and longitudinal coordinates of location X, Y where each camera image was taken. It also stores timestamp T of data logging event synchronized with the video stream. Therefore our first automated step is to convert GPS coordinates X, Y, T from data logger to dA, dV (delta angle and delta speed) for each T. By predicting dA and dV instead of X and Y we could directly provide input to vehicle control unit to actuate driving wheel angle and accelerator. Since our machine learning system must be trained on finite number of classes, we decided to compute independently histograms of value distribution for dA and dV; then quantize them into 20 bins each. The resulting resolution of single bin for dA was ~1.15degree and dV was ~0.05G. Quantization was performed in such a way to keep 95% of measurements within 20 bins for both dA and dV, while all values outside of 95% were aligned to the closest value within the range +-11.5 degree for dA and +-0.5G for dV. Finally, since both angle and speed needs to be accurate to compute future vehicle position we made totally 20x20=400 classes of labels (corresponds to 20bins angle x 20bins speed). Given the size of our dataset we wanted to ensure that at least 1000 samples would fall into each category for proper training of deep neural networks. Each camera image would be further associated to one of 400 classes both for training and validation stages using the same procedure described earlier.

Database to train and validate CNN was constructed in the following way:

- Slide time-window of size 2 sec (past) + 1 sec (future) over recorded videos (with synchronized GPS)
- Fetch images from T-2 to T for each time-window
- Make combined image (space-time) from T-2 (R), T-1(G) and T(B) as RGB equivalent image, use it as an input with dA, dV computed at T+1 and converted to 1:400 classes as output or labels (for the training).
- Augment data by randomly cropping camera image 32 times (50%-100% of image height), then convert each crop to 64x64 images and keep original label

We decided to split data into training 50% and test 50% sets, in such a way that each road was driven at least two times, so that at least one driving scenario is available for training and similar driving scenario (potentially driven another day by other driver) is available for testing.

Architecture of Convolutional Neural Network used for Control Point Prediction

The convolutional neural network used in this paper was inspired by [3]. Such architecture was proven to be capable to find good visual features for wide range of everyday visual inputs and to be able to discriminate between visually similar classes of objects. We consider this architecture as a baseline version of a neural network that is deep enough to handle everyday driving situations and to discriminate between visually similar driving scenarios. It has 7 layers of hidden neurons. It uses local receptive fields of size 7×7, and a stride length of 2 pixels. There are a total of 96 feature maps. The first 5 hidden layers are convolutional layers (some with max-pooling), while the next 2 layers are fully-connected layers. The output layer was adapted. It is a 400-unit softmax layer, corresponding to the 400 image classes.

Data Acquisition for Control Point Prediction by CNN

In this section, we introduce the IMRA dataset of 1000 hours of driving videos (should be completed on March 2017). Driving videos along with GPS and accelerometer (3-axis) are recorded using aftermarket driving recording camera (also called Dash-cam) mini0806 at 30fps with resolution of 2.5Mpx. To represent drivers' view point, the dash-cam is installed at the vehicle's center near rearview mirror. Overall eight cars are used to record driving videos of everyday situations in France and Germany. In each country driving zone is in 90% of cases limited to 50x50 km region. One or several drivers were driving each of 8 cars.

Evaluation of Single Control Point Prediction by CNN

Experimental evaluation shows that CNN could predict future changes of wheel angle and acceleration required for safe single control point (1sec prediction) with 95% accuracy in case of a dataset limited to 10 hours of everyday driving scenarios. Accuracy of 95% for our system is computed as 100*M/N, where M=95 is the number of trials for which dA and dV errors between predicted and reference data was less than 1.15 degree for angle and 0.05G for acceleration, given N=100 is the total number of trials. Six categories of driving scenarios, each represented by ten events of duration 30-180 seconds were analyzed. We could understand that given limited amount of data (10 hours), the CNN could learn to recognize typical driving situations mostly related to environmental scenes. It was able to accurately predict future position within 1 seconds by efficiently combining time-space image input and information how drivers behaved in typical visually similar situations. Table 1 shows the distribution of accuracy for the single control point prediction by CNN.

Scenario	1	2	3	4	5	6
	Roundabout	Road Bump	Sharp Turn	Sharp Brake	Obstacle	Suburb
Accuracy %	91.5	92.5	92.9	93.3	93.4	95.0

Table 1 - Accuracy of CNN single control point prediction for six categories of driving scenarios

However, in case of rare/unique events (scenario 1-5 in Table 1), we could observe decrease of prediction accuracy. One of the reasons for that is an absence of enough data samples that represent various dynamic scenes in case of visually similar environment, so that the deep neural network couldn’t find enough high level features to discriminate between these situations without error.

Next section will discuss how to extend the 1 sec interval and improve prediction accuracy, by introducing larger training datasets and multiple control point prediction (7 seconds ahead) using similar deep learning architecture.

Deep Learning based Multiple Control Point Prediction

In the previous section, we could demonstrate that in limited conditions a deep neural network could be trained to predict the required control point of a car within 1 second using camera images as the only input. In the following section, we describe how to adapt such a system to predict multiple control point-based own vehicle trajectory up to 7 seconds ahead. We decided to set 7 second limit as a practical observation which corresponds to the limits of visible road area for combined urban/suburb driving. We have found that in more than 50% of driving scenarios in our database, a human is unable to predict future vehicle trajectory longer than for 7 seconds. We also introduced a new measure to evaluate of how deep learning based trajectory prediction is accurate compared to the reference data. The GPS data from the logger was interpolated using Kalman filtering algorithm, so that mean relative localization accuracy is between 1-2m error along 100m distance. We decided to fix the target error limit of trajectory prediction to 10% of the distance driven within 7 second time interval, which is the upper bound of practical driver assistance applications [2].

We consider few practical applications of such end-to-end vehicle trajectory prediction system. First, we consider a driver assistance function, where our system is employed as a co-pilot. Such co-pilot constantly monitors road situations and makes predictions based on deep learning algorithm described in this paper where the actual driver could be up to 7 seconds ahead. At the same time, actual driving trajectory is compared with co-pilot’s past prediction and in case of strong disagreement between the two an alert in form of audio/vibration or visual signal has to be issued. Another application is to evaluate how current driver compares to the aggregated driving experience learned by the co-pilot. It could be used to asses it’s driving skill, identify those driving situations where a human was not safe and use this as an education material. Last application is related to autonomous driving, where such system being trained on very large scale of data could be used to actuate vehicle control.

Data Collection and Processing for Multiple Control Point Prediction

In the previous sections, we described how we converted real values of dA, dV into 1:400 labels used for prediction of single control point. Although, the discretization accuracy was sufficient for a single prediction, the problem of multiple control point prediction is much more complex (10^18) if defined in similar fashion.

Therefore, we decided to cluster each 7 second trajectory composed of 7 control points (each point corresponds to 1:7 second of future position with respect to the current position) into predefined number of categories, so that we could keep number of labels for deep learning algorithm below practical limit (<10^5). As a result of the clustering method, each reference 7 control point-based trajectories computed from ground truth GPS data were converted into 1024 categories. Figure 4 illustrates example results of this procedure.

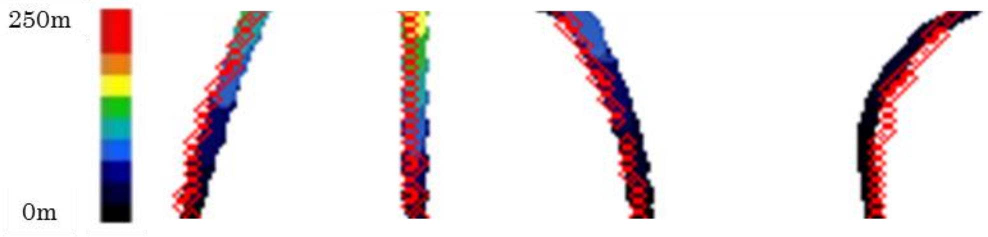


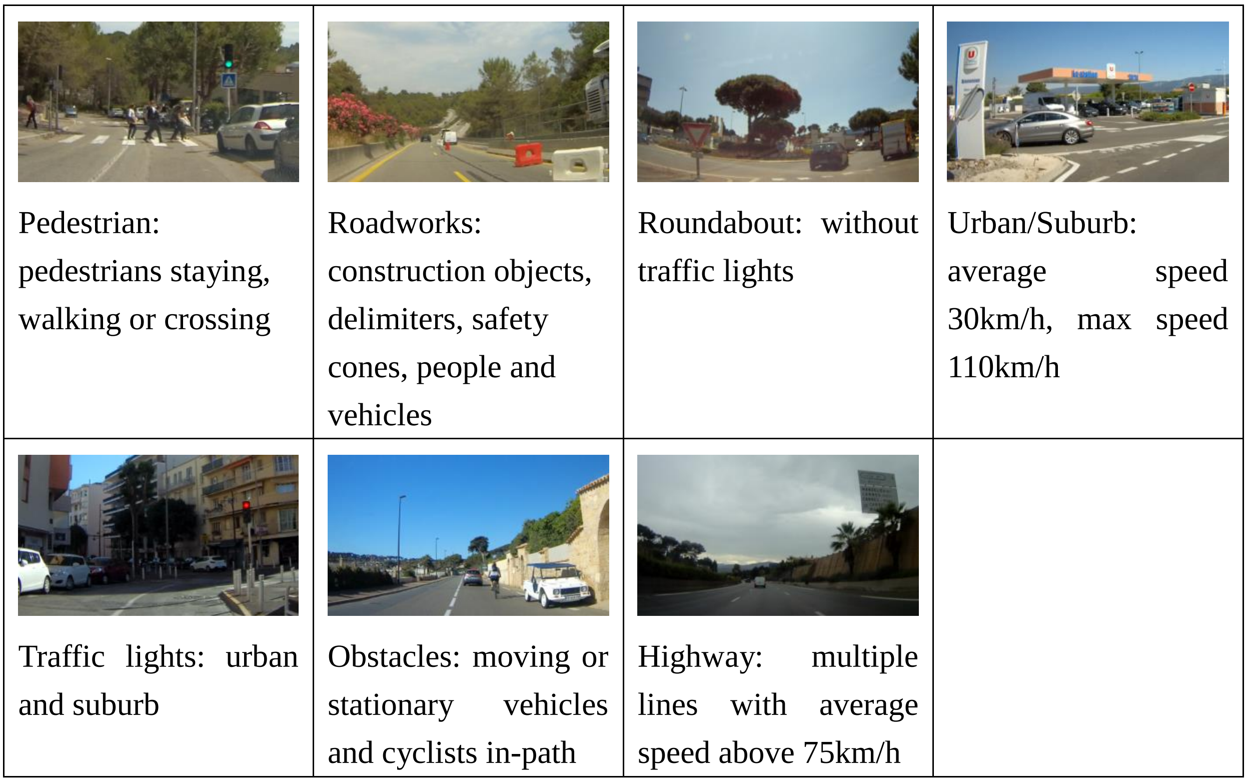
Figure 4 – Clustering of 7 control point trajectories. Four samples of real trajectory (red circles) and the corresponding category found by our clustering process. Each cluster is defined as color coded superposition of trajectories, where color is associated with future travelled distance (see legend on the left side)

Two types of measures could be used to evaluate quality of trajectory prediction by CNN. First, we could measure how many of the control points (out of total number of predicted control points) are predicted within 10% or less error compared to the reference distance. Averaged for a time period of the test data, this information is useful to estimate the probability that the correct prediction (within target accuracy range) would be available for particular application of interest at any time moment. This measure is named “in Target” in the next section. Second, we could measure overall error of prediction vs. reference distance for all control points, which is related to the overall accuracy of prediction. This measure is named “Error” in the next section.

Similar process as described in the section of single control point data processing was used to generate image data for both training and test of CNN. Three major differences are related to the following aspects: additional use of speed information by CNN at both training and test stage (to reduce overfitting of CNN), size of output layer of CNN equal to 1024 categories and use of 42 hours of data driven by 8 drivers for training and 4 hours for testing, where training and test are mostly happen in different geographical area or separated by significant time interval, longer than several months.

Evaluation of Multiple Control Point Prediction by CNN

Seven driving scenarios, each one coming from 15 different locations, were evaluated during the test. Totally all scenarios are represented by 4 hours of driving recordings, where urban/suburb scenario being 3.5h long and others are 5 minutes each totally 30 min long. Figure 5 shows sample key frames for each scenario.



Pedestrian: pedestrians staying, walking or crossing

Roadworks: construction objects, delimiters, safety cones, people and vehicles

Roundabout: without traffic lights

Urban/Suburb: average speed 30km/h, max speed 110km/h

Traffic lights: urban and suburb

Obstacles: moving or stationary vehicles and cyclists in-path

Highway: multiple lines with average speed above 75km/h

Figure 5 – Keyframes of driving scenarios used for evaluation of multiple control point prediction using CNN

The results of seven control-points prediction in case of trajectory clustering with 1024 clusters are summarized in table 2.

Scenario	CNN Error (% of distance)	CNN in Target (%) 1-7 sec
Pedestrian	22.3	51.2
Roadworks	12	49.5
Roundabout	9.6	74.6
Traffic Light	16.2	58.8
Obstacles	12	55.2
Highway	1	95.5
Urban/Suburb	5.0	98.6

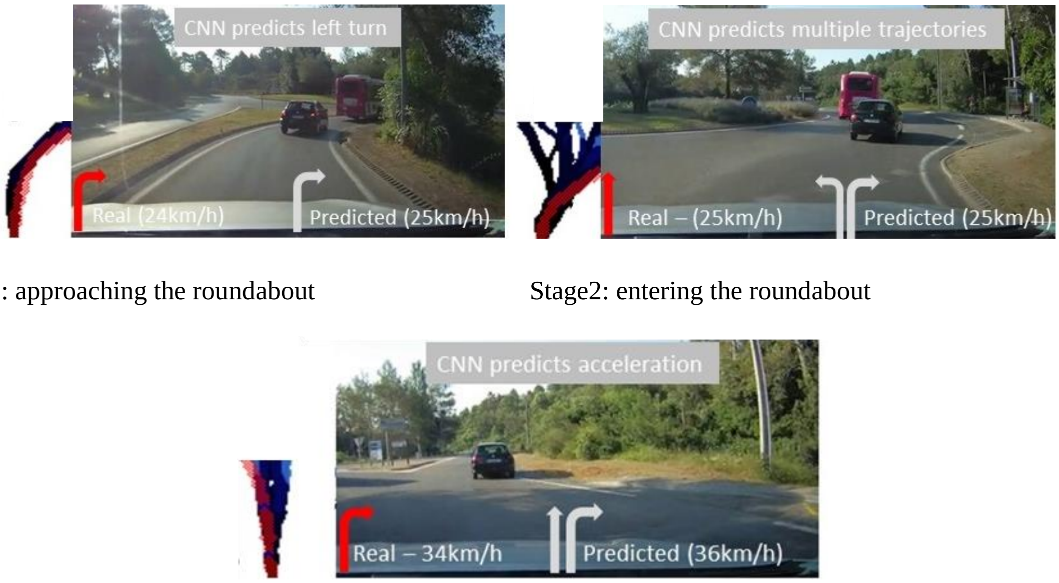
Table 2 - Distribution of errors of CNN multiple control point prediction for seven types of driving scenarios

One could see that in case of 1024 trajectory categories, our CNN based 7 control point prediction system was able to match the reference ground truth data in case of urban/suburb (5% error of distance) and highway (1% error of distance) driving scenario. Table 3 presents results of how absolute prediction error varies with distance, which turns to be almost distance-independent with value of ~1.6m.

Distance (m)	Error (% of distance)	Error (m)
0-20	8.4	1.68
20-40	4.1	1.64
40-60	2.7	1.62
60-80	2.2	1.76
80-100	1.5	1.5
100-120	1.3	1.5

Table 3 –CNN multiple control point prediction error vs. distance

This is an important result since the training/test was done on videos separated geographically or with several month time intervals between acquisitions. It shows the ability of this end to end CNN approach to integrate experience of multiple drivers and have successful trajectory prediction in previously unseen situation. Figure 6 illustrates how experience of multiple drivers was merged by CNN into single prediction. This example shows the trajectory prediction and speed predictions in different stages when driving in the roundabout scenario.



Stage1: approaching the roundabout

Stage2: entering the roundabout

Stage3: going out of the roundabout

Figure 6 – Multiple control point prediction using CNN for roundabout scenario

It could be seen how CNN, which was only trained using a single trajectory at time example, was able to correctly predict not only right turn before entering the roundabout but also multiple possible ways to drive through it. It also correctly suggested accelerating at the exit.

Overall two scenario (urban/suburb and highway) provide predictions within target range of 10% in >95% of time.

While being partially solved, other driving scenarios happen to be much rarer compared to the typical driving situations. Achieving similar accuracy in these cases will require improvement of the proposed system by augmenting data and increasing deep learning generalization capacity in case of rare/unique events.

Conclusion

In this paper, we have presented the first system that predicts future trajectory using only video images as input and deep neural networks for decision making. We demonstrated that the convolutional neural network trained on 10 hours of driving data could predict with 95% accuracy required changes of wheel angle and acceleration within 1 second in future.

We have also shown that the same type of neural network trained on 42 hours of driving could predict future trajectory of vehicle 7 seconds ahead with 5% error of distance in 98.6% of time for typical urban/suburb driving scenario.

Thus we could conclude that end-to-end vehicle control system that doesn’t explicitly rely on obstacle detection could be realized by integrating experience of multiple drivers in visually similar situation using deep learning. In addition, we have shown the using only images, large time horizon prediction can be achieved. This long term prediction will certainly be a key to enable next generation of safety applications.

We also discussed how high accuracy of predicted information demonstrated in this paper could lead to driving assistance or automated vehicle control applications. We have identified two scenario (highway and urban/suburb) with highest trajectory prediction accuracy.

Future extension of this work would continue in three directions. First, we would train the system with more data; try different deep learning network architectures and data augmentation. Second, since the 7 second ahead prediction was achieved using for training the CNN only 2sec of previous video and GPS history, we should find the optimum length of history required for each scenario. This should lead to increase to prediction accuracy. Third, we would focus on improving deep learning capacity to avoid safety critical errors in case of rare/unique situations, which couldn't be seen during training stage but must be safely predicted during the online driving support.

References

1. SafetyNet (2009) Pedestrians & Cyclists, European Commission
2. Tsishkou, D. et al., "Real-time Multiple Object Recognition for Collision Avoidance Using Wide Angle Stereo Camera", 21st ITS World Congress, USA, 2014
3. Krizhevsky, A., et al., "ImageNet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing 25, MIT Press, Cambridge, 2012
4. RU2013102888, Method And Module For Vehicle Speed Control, 2014
5. JP2004268819, Steering Control Device And Steering Control Method Of Vehicle, 2004
6. GB251016, Vehicle Path Prediction And Obstacle Indication System, 2014
7. US2008088424, Active Safety Apparatus, 2010
8. WO2006077182, Driver Assistance System With Driving Path Prediction, 2006
9. US2010318240, Course Evaluation Apparatus And Course Evaluation Method, 2010
10. Jayaraman D. and Grauman K., "Learning image representations tied to ego-motion", ICCV 2015
11. Shashua A., What goes into sensing for autonomous driving? https://www.youtube.com/watch?v=GCMXXXmxG-I, 2016