

Two-Stream Networks for Lane-Change Prediction of Surrounding Vehicles

David Fernández-Llorca¹, Mahdi Biparva², Rubén Izquierdo-Gonzalo¹ and John K. Tsotsos²

Abstract—In highway scenarios, an alert human driver will typically anticipate early cut-in and cut-out maneuvers of surrounding vehicles using only visual cues. An automated system must anticipate these situations at an early stage too, to increase the safety and the efficiency of its performance. To deal with lane-change recognition and prediction of surrounding vehicles, we pose the problem as an action recognition/prediction problem by stacking visual cues from video cameras. Two video action recognition approaches are analyzed: two-stream convolutional networks and spatiotemporal multiplier networks. Different sizes of the regions around the vehicles are analyzed, evaluating the importance of the interaction between vehicles and the context information in the performance. In addition, different prediction horizons are evaluated. The obtained results demonstrate the potential of these methodologies to serve as robust predictors of future lane-changes of surrounding vehicles in time horizons between 1 and 2 seconds.

I. INTRODUCTION

One of the closest and most plausible scenarios in the adoption of the autonomous vehicles is autonomous navigation at SAE L3 (chauffeur) or L4 (autopilot) on highways, both for passenger and freight transport. The most advanced automation systems to date are the Highway Chauffeur (HC) and the Highway Autopilot (HA), which includes the management of complex maneuvers such as deciding to change lanes to overtake, enter a slower lane or even exit the highway. HC is mostly considered as L3 and HA as L4[1]. In these systems, the most critical, and challenging, highway scenarios are the cut-in and cut-out ones, specially for high speeds. In the cut-in scenario, a car from one of the adjacent lanes merges into the lane just in front of the ego car. In the cut-out scenario, a car in front leaves the lane abruptly to avoid a slower vehicle, or even stopped, ahead. Since 2018, the performance of these assistance or chauffeur commercial systems operating under these two critical traffic scenarios is being tested by Euro NCAP [2].

An alert driver will typically anticipate early cut-in and cut-out maneuvers using only visual cues, reduce speed accordingly or even change lanes through the use of the steering wheel. An automated system must also be able to anticipate these situations at an early stage. To do so, it is necessary to endow new automated systems with the ability of predicting the motions of surrounding vehicles, such as lane-keeping and lane-change.

¹D. F. Llorca and R. Izquierdo are with the Computer Engineering Department, University of Alcalá, Alcalá de Henares, Madrid, Spain david.fernandezl@uah.es

²M. Biparva and J. K. Tsotsos are with the Department of Electrical Engineering and Computer Science, York University, Toronto, ON M3J 1P3, Canada tsotsos@eecs.yorku.ca

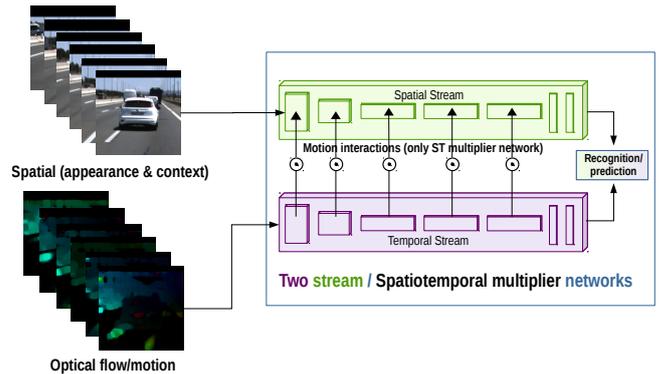


Fig. 1. Overview of the proposed video action recognition approaches for lane change recognition and prediction of surrounding vehicles, including Two-Stream Network and Spatiotemporal Multiplier Network.

To deal with lane-change prediction of surrounding vehicles, in this paper we pose the problem as an action recognition problem using visual information from cameras. The idea behind our proposal is to use the same source of information (visual cues) and the same type of approach (action recognition) that drivers use to anticipate these maneuvers.

Significant progress has been made in video-based human action recognition and prediction during the last years [3]. Action recognition and prediction involves managing spatial and temporal information (sequence of images). Among the different methodologies, in this paper, we study **Two-Stream Convolutional Networks** [4] and **Spatiotemporal Multiplier Networks** [5] approaches (see Figure 1) using The PREVENTION dataset [6] to train and validate them.

II. RELATED WORK

Most of the available work on lane-change recognition and prediction focuses on lane-changes of the ego-vehicle. However, the nature of the problem is considerably different when we focus on lane-changes of surrounding vehicles, so we limit our analysis of lane-change detection of other vehicles, within the context of the highway scenario.

A. Input variables

Most of the previous works are based on the use of physical variables that define the relative dynamics of the vehicle with other vehicles and with its environment [7], [8], [9], [10], [11], [12], [13],[14], [15], [16], [17], [18], [19], [20], [21], including lateral and longitudinal distances, velocity, acceleration, timegap, heading angle and yaw rate.

Context cues are also introduced, including road-level features such as the curvature and speedlimit [9], [11], [20],

distance to the next highway junction [13], number of lanes [19], etc., as well as lane-level features such as type of lane marking or the distance to lane end [13].

The number of proposals making use of appearance features is surprisingly low, especially considering that human drivers do not use the physical variables mentioned above to anticipate lane changes from other vehicles but visual cues. In [20], two variables manually selected from the appearance, i.e., state of turn indicators and state of brake indicators, are used. In our previous work [22] regions of interest (ROIs) are generated for each vehicle detection, including local information around the vehicle, and appearance features are extracted using a GoogLeNet pre-trained on ImageNet.

B. Methodologies

As suggested by [23] vehicle motion modeling and prediction approaches can be classified into three different levels: physical-based, where predictions only depend on the laws of physics, maneuver-based, where the future motion of a vehicle depends on the driver maneuver, and intention-aware, where predictions take into consideration inter-dependencies between vehicles.

Some proposals are intention-aware in their nature. For example, by using graphical models such as Bayesian Networks [7], [13], [20] or Structural Recurrent Neural Networks [19], or by using convolutional social pooling in an LSTM encoder-decoder architecture [18]. However, in most cases, inter-dependencies between vehicles are modeled by extracting relative physical features [9], [11], [15], [17] or by generating compact representations that encode the relative positions of all vehicles on the scene [16], [22]. Some works do not take into consideration the interaction between vehicles [8], [10], [12], [24], [21].

Many approaches to lane-change recognition and prediction address the problem using generative-based solutions, including Naïve Bayes Classifiers [9], Bayesian Networks [7], [13], [20], and Hidden Markov Models [10]. Others make use of discriminative solutions such as case-based reasoning [8], Random Decision Forest [11], traditional Neural Networks [12], [14], Support Vector Machines [14], [15], [24], Gaussian Process Neural Networks [21], and feed-forward Convolutional Neural Networks [16], [22]. Finally, some other approaches are based on the use of Recurrent Networks including vanilla LSTM [22] and LSTM encoder-decoder [17] and multi-modal [18] architectures.

C. Datasets

Two type of recording setups are usually proposed depending on the location of the sensors. First, we have datasets captured from the infrastructure using cameras installed on buildings, such as NGSIM HW101 [25] or NGSIM I-80 [26] datasets, or cameras on-board drones, such as HighD [27], inD [28] or INTERACTION [29] datasets. Although these datasets are very valuable for understanding and assessing the motion and behavior of vehicles and drivers under different traffic scenarios, they are not fully applicable for on-board detection applications.



Fig. 2. ROI sizes. From upper row to lower row: x_1 , x_2 , x_3 and x_4 . The vehicle is always centered. Zero-padding is applied when needed.

Second, other datasets provide road data with sensors on-board vehicles. In this line, the PKU dataset [30] contains 170 minutes of data gathered using a vehicle equipped with 4 2D-LiDARs covering a region of 40 meters around the vehicle (road lane markings, number of road lanes, or the relative positioning of the ego-vehicle are not provided). The ApolloScape dataset [31] provides data obtained in urban environments from 4 cameras and 2 Laser scanners using a vehicle driving at 30 km/h. It does not contain radar data, making detections more sensitive to failure in adverse weather conditions, and it does not provide labeled tracking information (IDs and tracklets) for all detected objects. Recently, in 2019, the PREVENTION dataset [6] was released containing data from 3 radars, 2 cameras and 1 LiDAR, covering a range of up to 80 meters around the ego-vehicle. Road lane markings are included and the final position of the vehicles is provided by fusing data from the three type of sensors.

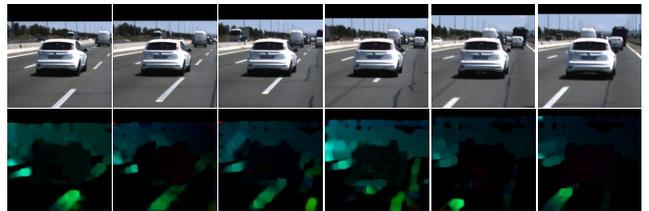


Fig. 3. Example of dense optical flow computation.

III. PROBLEM FORMULATION

We define lane change prediction as a multi-classification problem in which the goal is to recognize whether a vehicle i will make a left or right lane-change or remain in its lane given the observed context up to some time N . The prediction relies on visual cues that are computed from regions of interest (ROIs) extracted from the contour labels provided in the PREVENTION dataset. Four different ROI sizes are considered: $\times 1$, $\times 2$, $\times 3$ and $\times 4$ the size of the

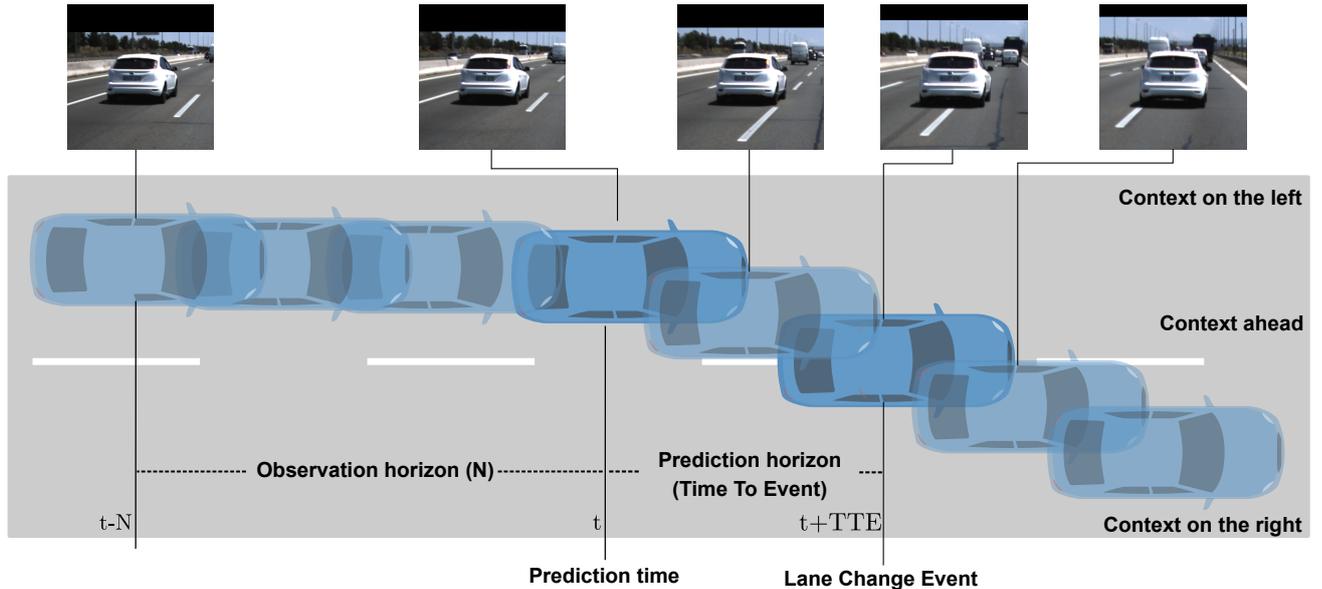


Fig. 4. Problem formulation: observation horizon (N), and time to event (TTE). The lane change event is labeled as the frame where the middle of the rear bumper is located just over the lane markings. This is the criterion established in PREVENTION dataset [6].

square bounding box around the vehicle contour (see Figure 2). Zero-padding is used when the ROI exceeds the limits of the image. The size of the ROI modulates the amount of context information being considered in the input data stream. Thus, $\times 1$ mostly contains information related with the vehicle appearance, while $\times 4$ incorporates a large amount of front and side context information. For ROI sizes greater than $\times 2$, the approach can be considered as interaction-aware.

Since the vehicle is always centered in the ROI, dense optical flow (from the motion stream) should be interpreted as a way of measuring the movement of the context (infrastructure and other vehicles) around the detected vehicle. As shown in Figure 3, the optical flow is low in the region where the vehicle is, while it is more predominant around it.

As can be seen in Figure 4, the lane-change event is defined as the time when the center of the rear bumper is just above the lane markings. The observation horizon or time window will contain a set of N images that will be stacked according to the activity recognition method used. We will examine the effects of time to event (prediction horizon) and observation duration (N) on the accuracy of lane-change classification (when $TTE = 0$) and prediction (when $TTE > 0$).

IV. VIDEO ACTIVITY RECOGNITION AND PREDICTION

The sequence of stacked images or regions of interests, can naturally be decomposed into spatial and temporal components. The spatial part, in the form of individual region appearance, carries information about the vehicle itself (e.g., light indicators or brake lights) and the context around it (road, lane markings and surrounding vehicles). The temporal part, in the form of motion across frames, conveys the movement of the observer (onboard camera)

w.r.t. to the road, and the surrounding vehicles. In order to handle a canonical view for the motion stream, all the regions are generated around the contour of the vehicle so the vehicle is always centered in the region of interest (the size will vary depending on the relative distance w.r.t. the ego vehicle). We consider two video activity recognition approaches: Disjoint Two-Stream Convolutional Networks [4] and Spatiotemporal Multiplier Networks [5].

A. Disjoint Two-Stream Convolutional Networks

A two-stream ConvNet architecture which incorporates and fuses spatial and temporal information is defined. The structure of the ConvNets for both streams is the same, including 5 convolutional layers and 3 fully connected layers, with the parameters depicted in Figure 5. The last fully connected layer is defined with 3 outputs regarding the three classes defined: left lane change (LLC), right lane change (RLC), and no lane change (NLC).

The dense optical flow is computed using polynomial expansion [32]. The spatial stream ConvNet is pre-trained using ImageNet and the temporal ConvNet using multi-task learning using UCF-101 and HMDB-51. All hidden layers use the rectification (ReLU) activation function. Max-pooling is performed over 3×3 spatial windows with stride 2.

B. Spatiotemporal Multiplier Networks

The original two-stream architecture only allows the two processing streams (spatial and motion) to interact via late fusion of their respective softmax predictions. This way, the architecture does not support the learning of truly spatiotemporal features, since the loss of both streams is back-propagated independently without any type of interaction. Learning spatiotemporal features requires the appearance and motion paths to interact earlier on during the forward

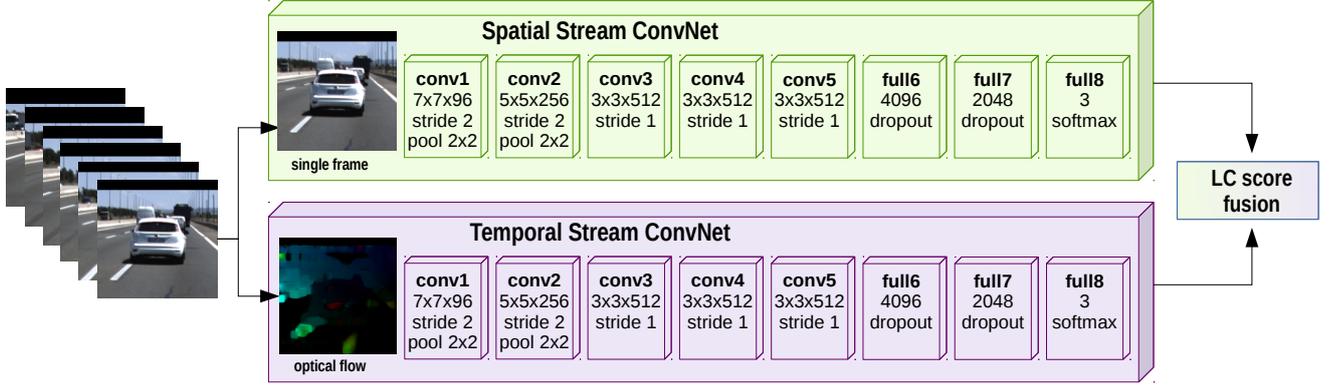


Fig. 5. Disjoint two-stream architecture for lane change classification and prediction.

pass. This interaction can be relevant for the classification and prediction of lane change maneuvers that have similar appearance or motion patterns and can only be inferred by the combination of two (e.g., vehicles that do not change lanes but have their turn indicators on). To address this limitation, it is possible to inject cross-stream residual connections using Residual Networks (ResNets) [33] as the general architecture for the spatial and the temporal streams.

In [5], different cross-stream connections were studied, including two types of connections (direct or into residual units), two fusion functions (additive or multiplicative), and different streams directions (unidirectional from the motion into the appearance, conversely and bidirectional), being the multiplicative residual connection from the motion path into the appearance stream the one providing the superior performance.

As can be observed in Figure 6, the multiplicative interaction can be formulated as:

$$\hat{x}_{l+1}^a = f(x_l^a) + \mathcal{F}(x_l^a \odot f(x_l^m), W_l^a) \quad (1)$$

where x_l^a and x_l^m are the inputs of the l -th layers of the appearance and motion paths respectively, while W_l^a represents the weights of the l -th layer residual unit in the appearance stream and \odot corresponds to elementwise multiplication.

Better temporal support is also provided by injecting 1D temporal convolutions layers into the network [5]. ResNet50 model is used for both streams, including batch normalization and ReLU activation function after each convolutional block.

C. Recognition & Prediction

The proposed two-stream architectures have been historically applied to perform activity recognition from video sequences (e.g., human activity recognition), i.e., using a sequence of images (from $t - N$ frames up to t) to perform the recognition of the activity taking place in the video at time t . In order to perform prediction (at $t + TTE$), we define the target class (none, left or right) that will take place in a future time horizon given by TTE as the desired

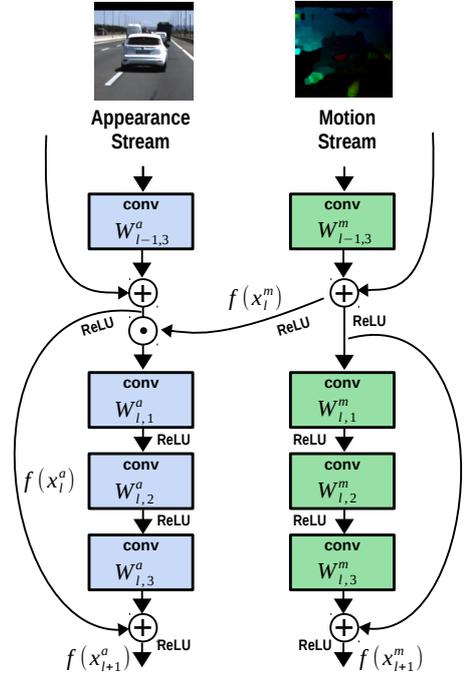


Fig. 6. Multiplicative residual gating from the motion stream to the appearance stream.

class at t . In other words, as can be observed in Figure 4, we can consider that the system is performing lane change recognition or classification at time t when $TTE = 0$. In the same way, the system will be predicting the lane changes when $TTE > 0$ (in frames). Taking into account that the sampling frequency of PREVENTION dataset is 10Hz, we define the following scenarios:

- $TTE = 0$ frames: lane-change classification at time t .
- $TTE = 10$ frames: lane-change prediction 1 second ahead ($t + 10$).
- $TTE = 20$ frames: lane-change prediction 2 seconds ahead ($t + 20$).

V. EXPERIMENTAL RESULTS

A. Dataset description

Table I summarizes the details of the dataset. The input size for both streams is 112×112 . The 85% of the samples are used for training and the remaining 15% for validation.

	NLC	LLC	RLC
# of sequences	3110	342	438
avg. # of frames	50.9	96.8	80.1

TABLE I

MAIN STATS OF THE DATASET. NLC/LLC/RLC: NO/LEFT/RIGHT LANE-CHANGE.

B. Evaluation parameters

The following parameters have been evaluated during the experiments:

- ROI sizes: x1, x2, x3 and x4.
- Observation horizon: 20 frames (2 seconds), 30 frames (3 seconds) and 40 frames (4 seconds).
- Time-to-event (prediction horizon): 0 (no prediction), 10 (1 second) and 20 (2 seconds).

C. Metrics

As a multi-class problem (with 3 classes), we have considered the accuracy as the main variable to assess the performance of the two evaluated methods and the corresponding parameters, i.e., the number of true positives for the three classes divided by the total number of samples.

D. Lane change classification results

In Table II we depict the accuracy of the two action recognition approaches over the validation set, i.e., with $TTE = 0$. Regarding the ROI sizes we can state the following conclusions. By using just the ROI fitted to the bounding box, the results are surprisingly reasonable, considering that almost no context and interaction are available. In general, the higher the ROI size, the better the accuracy. However for observation horizon of 40 frames, adding more context from x3 to x4 decreases the performance. This can be explained by the fact that a larger observation horizon already incorporates more context into the spatial and motion streams.

For observation horizons of 20 and 30 frames, the simpler disjoint two-stream network offers better results than the spatiotemporal multiplier network. However, for larger observation horizons (40 frames), the added complexity of the cross-stream residual connections yields the best performance, i.e., an accuracy of 90.3% (see Table II). Note that these results clearly outperform previous results on the PREVENTION dataset [22].

E. Lane change prediction results

The ability of both methodologies to predict the future lane-change manoeuvre of surrounding vehicles is evaluated using an observation horizon of 20 frames (2 seconds) and prediction horizon (TTE) of 10 and 20 frames. The results for both approaches are depicted in Table III.

Method	Obs. Horizon	ROI size			
		x1	x2	x3	x4
Disjoint	20	83.22	86.18	86.26	87.43
Disjoint	30	83.55	86.69	86.84	86.68
Disjoint	40	84.97	87.69	89.46	88.79
ST	20	83.39	85.03	86.51	86.16
ST	30	84.38	84.70	85.36	84.73
ST	40	86.02	87.83	90.30	89.64

TABLE II

DISJOINT TWO-STREAM NETWORK AND SPATIOTEMPORAL MULTIPLIER NETWORK CLASSIFICATION ACCURACY (%).

As can be observed, surprisingly, the results for longer prediction horizons are better, i.e., the accuracy of both models for $TTE = 20$ frames is approximately 5% higher in all cases for both models than for a $TTE = 10$. This can be partially explained by the complexity of the two-stream models that improves generalization with a more complex target to learn. The best accuracy for the disjoint two-stream network is given for a prediction horizon of 2 seconds and a ROI size of x3, yielding 91.02%. For the spatiotemporal multiplier network, the best prediction accuracy, 91.94% is given for a TTE of 2 seconds and a ROI size of x4.

Up to our knowledge, these are the first prediction results so far using the PREVENTION dataset.

Method	TTE	ROI size			
		x1	x2	x3	x4
Disjoint	10	84.05	84.54	85.20	85.36
Disjoint	20	85.20	88.82	91.02	90.92
ST	10	84.70	85.69	85.20	86.51
ST	20	86.84	90.30	91.45	91.94

TABLE III

DISJOINT TWO-STREAM NETWORK AND SPATIOTEMPORAL MULTIPLIER NETWORK PREDICTION ACCURACY (%). OBSERVATION HORIZON = 20.

VI. CONCLUSIONS

In this paper, two video action recognition approaches have been adapted, trained and evaluated to perform lane-change classification and prediction of surrounding vehicles in highway scenarios using the PREVENTION dataset. The problem was posed as an action recognition problem using visual cues from cameras, i.e., using the same source of information and approach that human drivers use to anticipate these maneuvers.

Both approaches, the disjoint two-stream convolutional network and the spatiotemporal multiplier network, are based on two different paths obtained from the same sequence of frames: a spatial stream in the form of individual region appearance, and a motion stream in the form of dense optical flow across frames. The complexity of the second model is based on the use of more complex architectures (ResNet50) and the use of multiplicative residual gating from the motion stream to the appearance stream.

Different ROI sizes have been evaluated, being the larger regions (x3 and x4) the ones providing the better results, due

to the fact that they implicitly incorporate context information and iteration with other vehicles. The best configuration (spatiotemporal multiplier network with ROI size of $\times 4$ and observation horizon of 2 seconds) yields almost a 92% of lane-change prediction accuracy two seconds earlier.

As future works, we plan to evaluate other action recognition approaches such as I3D models [34] and SlowFast Networks [35].

ACKNOWLEDGMENT

This work was supported in part by Spanish Ministry of Science, Innovation and Universities (Salvador de Madariaga Mobility Grant PRX18/00155 and Research Grant DPI2017-90035-R) in part by the Community Region of Madrid (Research Grant 2018/EMT-4362 SEGVAUTO 4.0-CM) and in part by the Air Force Office of Scientific Research USA (FA9550-18-1-0054) the Canada Research Chairs Program (950-231659) and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-05352).

REFERENCES

- [1] ERTRAC, "Connected automated driving roadmap," 8 of March 2019, site: <https://www.ertrac.org/uploads/documentssearch/id57/ERTRAC-CAD-Roadmap-2019.pdf>.
- [2] EuroNCAP, "2018 automated driving tests," October 2018, site: <https://www.euroncap.com/en/vehicle-safety/safety-campaigns/2018-automated-driving-tests/>.
- [3] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," 2018, arXiv preprint arXiv:1806.11230.
- [4] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014, pp. 568–576.
- [5] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4768–4777.
- [6] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, "The prevention dataset: a novel benchmark for prediction of vehicles intentions," in *IEEE 22nd International Conference on Intelligent Transportation Systems (ITSC)*, 2019, pp. 3114–3121.
- [7] D. Kasper, G. Weidl, T. Dang, G. Breuel, A. Tamke, A. Wedel, and W. Rosenstiel, "Object-oriented bayesian networks for detection of lane change maneuvers," *IEEE Intelligent Transportation Systems Magazine*, vol. 4, no. 3, 2012.
- [8] R. Graf, H. Deusch, M. Fritzsche, and K. Dietmayer, "A learning concept for behavior prediction in traffic situations," in *IEEE Intelligent Vehicle Symposium (IVS)*, 2013, pp. 672–677.
- [9] J. Schlechtriemen, A. Wedel, J. Hillenbrand, G. Breuel, and K.-D. Kuhnert, "A lane change detection approach using feature ranking with maximized predictive power," in *IEEE Intelligent Vehicle Symposium (IVS)*, 2014, pp. 108–114.
- [10] P. Liu, A. Kurt, and U. Ozguner, "Trajectory prediction of a lane changing vehicle based on driver behavior estimation and classification," in *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, 2014, pp. 942–947.
- [11] J. Schlechtriemen, F. Wirthmueller, A. Wedel, G. Breuel, and K.-D. Kuhnert, "When will it change the lane? a probabilistic regression approach for rarely occurring events," in *IEEE Intelligent Vehicle Symposium (IVS)*, 2015, p. 13731379.
- [12] S. Yoon and D. Kum, "The multilayer perceptron approach to lateral motion prediction of surrounding vehicles for autonomous vehicles," in *IEEE Intelligent Vehicle Symposium (IVS)*, 2016, p. 13071312.
- [13] M. Bahram, C. Hubmann, A. Lawitzky, M. Aeberhard, and D. Wollherr, "A combined model- and learning-based framework for interaction-aware maneuver prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 6, p. 15381550, 2016.
- [14] R. Izquierdo, I. Parra, J. M. noz Bulnes, D. Fernández-Llorca, and M. A. Sotelo, "Vehicle trajectory and lane change prediction using ann and svm classifiers," in *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017.
- [15] W. Yao, Q. Zeng, Y. Lin, D. Xu, H. Zhao, F. Guillemard, S. Geronimi, and F. Aioun, "On-road vehicle trajectory collection and scene-based lane change analysis: Part ii," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, p. 2062220, 2017.
- [16] D. Lee, Y. P. Kwon, S. McMains, and J. K. Hedrick, "Convolution neural network-based lane change intention prediction of surrounding vehicles for acc," in *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017.
- [17] N. Deo and M. M. Trivedi, "Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms," in *IEEE Intelligent Vehicle Symposium (IVS)*, 2018, p. 11791184.
- [18] —, "Convolutional social pooling for vehicle trajectory prediction," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2018, p. 14681476.
- [19] S. Patel, B. Griffin, K. Kusano, and J. J. Corso, "Predicting future lane changes of other highway vehicles using rnn-based deep models," 2018.
- [20] J. Li, B. Dai, X. Li, X. Xu, and D. Liu, "A dynamic bayesian network for vehicle maneuver prediction in highway driving scenarios: Framework and verification," *Electronics*, vol. 8, no. 40, 2019.
- [21] M. Kruger, A. S. Novo, T. Nattermann, and T. Bertram, "Probabilistic lane change prediction using gaussian process neural networks," in *IEEE 22th International Conference on Intelligent Transportation Systems (ITSC)*, 2019, p. 36513656.
- [22] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, "Experimental validation of lane-change intention prediction-methodologies based on cnn and lstm," in *IEEE 22th International Conference on Intelligent Transportation Systems (ITSC)*, 2019, p. 36573662.
- [23] S. Lefevre, D. Vasquez, and C. Laugier, "A survey on motion prediction and riskassessment for intelligent vehicles," *ROBOMECH Journal*, vol. 1, no. 1, 2014.
- [24] J. Li, C. Lu, Y. Xu, Z. Zhang, J. Gong, and H. Di, "Manifold learning for lane-changing behavior recognition in urban traffic," in *IEEE 22th International Conference on Intelligent Transportation Systems (ITSC)*, 2019, p. 36633668.
- [25] J. Colyar and J. Halkias, "Ngsim - us highway 101 dataset," 2007, site: <https://www.fhwa.dot.gov/publications/research/operations/07030/>.
- [26] J. Halkias and J. Colyar, "Ngsim - interstate 80 freeway dataset," 2006, site: <https://www.fhwa.dot.gov/publications/research/operations/06137/>.
- [27] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems," in *IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018.
- [28] J. Bock, R. Krajewski, T. Moers, L. Vater, S. Runde, and L. Eckstein, "The ind dataset: A drone dataset of naturalistic vehicle trajectories at german intersections," 2019.
- [29] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clause, M. Naumann, J. Kümmerle, H. Königshof, C. Stiller, A. de La Fortelle, and M. Tomizuka, "INTERACTION Dataset: An INTERNATIONAL, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps," *arXiv:1910.03088 [cs, eess]*, 2019.
- [30] H. Zhao, C. Wang, Y. Lin, F. Guillemard, S. Geronimi, and F. Aioun, "On-road vehicle trajectory collection and scene-based lane change analysis: Part i," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, p. 192205, 2017.
- [31] P. Wang, X. Huang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloscape open dataset for autonomous driving and its application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [32] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," *Lecture Notes in Computer Science*, vol. 2749, pp. 363–370, 2003.
- [33] S. K. He, X. Zhang and J. Sun, "Deep residual learning for image recognition," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [34] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetic dataset," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [35] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Quo vadis, action recognition? a new model and the kinetic dataset," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6202–6211.