A REVIEW OF DATA-DRIVEN LANE-CHANGING DECISION MODELING FOR CONNECTED AND AUTOMATED VEHICLES

Zhengwen Fan
Master candidate,
School of Automation,
Nanjing University of Science and Technology, Nanjing, China, 210094
matthew_fan@njust.edu.cn

Shanglu He
Associate Professor,
School of Automation,
Nanjing University of Science and Technology, Nanjing, China, 210094
she17@njust.edu.cn
Corresponding author

Xinya Zhang
Master candidate,
School of Automation,
Nanjing University of Science and Technology, Nanjing, China, 210094
zxy199903242021@163.com

Yingshun Liu
Associate Professor,
School of Automation,
Nanjing University of Science and Technology, Nanjing, China, 210094
yingshun@njust.edu.cn

ABSTRACT

Lane-changing is a critical driving behavior of connected and automated vehicles (CAVs). This research provided a state-of-the-art review of the data-driven lane-changing decision (LCD) modeling for CAVs. The first step was to perform a knowledge graph co-occurrence analysis on keywords associated with data-driven LCD models. Accordingly, the existing research was summarized from two perspectives. One is based on the used data sources. The extensively utilized data sources, their properties, the primarily used settings, and the applicable scenarios were all summarized in this study. The other perspective is based on LCD modeling methods. The prevalent modeling methods and the accompanying methodologies for validation and evaluation were covered. Based on these findings, three future research directions were
concluded for data-driven LCD models of CAVs, i.e., the demand for a more comprehensive dataset that includes the characteristics of drivers and the mixed flow environment, the novel data-driven methods, the unified test set, and test standards. The results of this study are expected to contribute to the development of more precise and effective LCD models for CAVs.

INTRODUCTION

Vehicle motion planning and decision-making are crucial elements of CAVs technology. Lane-changing (LC) behavior is one of the most common driving habits when cars are moving. LC refers to the driving behavior of leaving the current lane and merging into an adjacent lane to achieve a desired driving objective. After considering a number of traffic variables, such as the speeds and separation of nearby vehicles, the state of the roads, and traffic management, this behavior is carried out. The LC behavior must execute the lateral motion required for lane switching while also considering the influence on the following car and the car-following relationship between the leading car on the original lane and the target lane. As a result, the LC behavior process is more complex than that of car-following behavior, and the related research findings are relatively limited (1).

The modeling of LCD is an abstraction of the decision-making process in the context of lane-changing behavior, outlining the logical principles and decision-making process that driving systems (or drivers) employ to choose whether to change lanes. This modeling focuses on expressing the micro-level LCD process and calibrating physical parameters. With the arrival of the big data era, researchers can benefit from the rapidly advancing data collection technology, which allows them to use high-precision, large-sample vehicle motion data and employ theoretical methods such as machine learning and data science. The underlying principles governing vehicle lane-changing decision-making can be discovered by researchers through training, learning, and iterating on the sample data. Such data-driven approaches for LCD also offer practical and human-like CAV decision-making.

An investigation of the lane-changing behavior of human vehicles (HV) must serve as the foundation for any human-like LCD for CAVs. There has been much research in recent years that has created data-driven models for LCD. This paper examines data sources, data features, modeling methods, and verification in-depth with a focus on data-driven LCD models. The article includes an overview of the current state of the research and an outlook on potential development tendencies.

ANALYSIS OF EXISTING RELEVANT RESEARCH BASED ON KNOWLEDGE GRAPHS

This study gathered 385 Chinese and 248 English papers that were relevant to data-driven LCD models from the CNKI and Web of Science databases between 1998 and 2022 in order to acquire a thorough grasp of the primary research content and future research directions of data-driven LCD models.
Lane-changing models (131 times), car-following models (37 times), lane-changing decision (23 times), driving behavior (16 times), and vehicle-road collaboration (14 times) are the most frequently used keywords in the selected CNKI Chinese literature. The lane-changing decision unit's co-occurrence knowledge graph is shown in Figure 1. Figure 1 shows that there were 49 articles with the term "lane-changing decision" in them. Ten articles' keywords mention either cellular automata or intelligent transportation, while eight articles' keywords include lane-changing trajectory. Microsimulation, safe distance, and car-following models, which correlate to rule-based lane-changing models and the use of microsimulation to acquire data or validate models, first appeared among the keywords connected to the lane-changing decision earlier (about 2014). In recent years (after 2020), terms like autonomous driving, data-driven, and deep learning have become very closely associated with lane-changing decisions.

**Figure 1: Co-occurrence knowledge graph of the lane-changing decision unit in Chinese literature**

The most often occurring keywords in the literature from the chosen Web of Science Core Collection are "lane changing model," "data-driven," "autonomous vehicles," and "trajectory data," with 43, 26, 21, and 17 occurrences, respectively. As seen in Figure 2, "autonomous vehicles," "safety," "collision avoidance," and "reinforcement learning" all appeared simultaneously in the 26 articles that used the keyword "decision making," with 12, 8, 7, and 6 occurrences each. "Neural network," "car following," and "prediction" were the earlier appearing keywords in terms of publishing time. Deep reinforcement learning, automation, and naturalistic trajectory are some current terms that are strongly related to decision making.
In conclusion, the analysis of lane-changing behavior characteristics utilizing microscopic traffic data, such as vehicle trajectories in intelligent transportation systems, can be considered as the core emphasis of LCD research in recent years. To help CAVs achieve safer, more pleasant, and more effective autonomous driving, a precise lane-changing model can be built using the strong feature extraction capabilities of deep learning on a vast amount of historical data. As a result, the three following issues will be resolved: data sources and their characteristics, commonly used modeling methods and validation, and future research directions.

![Co-occurrence knowledge graph of the lane-changing decision unit in English literature](image)

**THE DATA SOURCES FOR DATA-DRIVEN LANE-CHANGING MODELS**

Table 1 provides an overview of the data sources used in the literature reviewed for this paper, their scenarios, the size of the samples that were collected, key fields, and other details.

Table 1 shows that the release of the high-precision vehicle trajectory data ([https://data.transportation.gov/Automobiles/Next-Generation-Simulation-NGSIM-Vehicle-Trajectory/](https://data.transportation.gov/Automobiles/Next-Generation-Simulation-NGSIM-Vehicle-Trajectory/)) by NGSIM (Next Generation Simulation) in 2007 can be interpreted as a turning point.
point in the study of data-driven LCD models. Prior to its introduction, researchers frequently used conventional techniques to collect data, including manual video calibration using cameras positioned at a height [33], on-road measurements with data collection vehicles [32], and interactive driving simulators [33]. Nevertheless, the data collected using these techniques fell short of the requirements of data-driven LCD models in terms of both quantity and accuracy. Since the release of NGSIM trajectory data, it has been the preferred choice for various research on data-driven lane-changing decision models. Nine articles, or 80% of the total, are represented in Table 1 as having used NGSIM as the data source for LCD models. Several modeling techniques have been investigated using the NGSIM data, including BP neural networks [4], support vector machines [6], and various deep learning-based LCD models [10].

It should be mentioned that NGSIM data is collected using cameras set up in a high location on the side of the road. Some researchers have found that these errors cannot be corrected through strict data cleaning or interpolation, especially at the junction of two camera frames where there is frequently a significant error in position and speed information due to the limitations of camera quality and image processing technology at that time. This erroneous data can create difficulties for data-based research on traffic flow theory. [34]

In recent years, with the advancements in drone technology and high-precision video acquisition, some open-source high-precision vehicle trajectory datasets have emerged. These include the German HighD dataset released in 2018, followed by the inD, rounD, and exiD datasets in subsequent years, the INTERACTION dataset released in 2019 containing multiple countries, and the Ubiquitous Traffic Eyes dataset, which was released by Southeast University in China.

The HighD dataset (https://www.highd-dataset.com) was collected by drones on six different highways near Cologne, Germany, and includes approximately 110,000 trajectories of cars and trucks. Compared to the NGSIM dataset, HighD has a larger sample size, more comprehensive microscopic traffic flow data, and higher data accuracy. Specifically, in terms of the number of recorded vehicles, the HighD dataset is nearly 12 times that of the NGSIM dataset. In terms of the recorded driving distance, the HighD dataset is nearly 9 times that of the NGSIM dataset. In addition, the proportion of trucks in the HighD dataset is 23%, significantly higher than the 3% in the NGSIM dataset, and there are relatively few trajectories in congested states [12]. The inD, rounD, and exiD datasets were built using the same approach for three different scenarios:
urban unsignalized intersections, roundabouts, and highway entrances and exits.

The INTERACTION dataset (http://interaction-dataset.com) is a comprehensive trajectory dataset published in 2019 that contains numerous collected scenarios from China, the United States, Germany, and Bulgaria, covering freeways, merging sections, roundabouts, and intersections. The large, varied dataset includes high-resolution maps with entire semantics, which can significantly reduce the amount of data preparation required.

The Ubiquitous Traffic Eyes dataset (http://seutraffic.com/#/home) is a drone video trajectory dataset released by Southeast University in China in 2020. It currently includes six maps, including expressways and the entrances and exits. The dataset uses complete video vehicle trajectory automatic extraction technology, which leads to a 0.1-second temporal resolution and a 0.01-meter position resolution.

In summary, for the past ten years, the NGSIM dataset has been the biggest measured microscopic traffic dataset made available to the research community. Its data forms the basis of the majority of data-driven studies on microscopic traffic flow, and numerous NGSIM-based model studies are frequently carried out using extensive research methods. Yet, in terms of data volume, accuracy, and the variety of traffic scenarios, recently published datasets have a better performance. Also, their appearance helps to solve the problems of the NGSIM dataset and also presents new opportunities and challenges for the validation and generalization testing of existing data-driven LCD models.
## Table 1: Commonly used datasets for data-driven lane-changing models

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Publishing Time</th>
<th>Literature</th>
<th>Data Collection</th>
<th>Country</th>
<th>Scenario</th>
<th>Number of Maps</th>
<th>Length of Collected Road Section(m)</th>
<th>Collection Hours</th>
<th>Driving Distance (km)</th>
<th>Number of Lanes per Direction</th>
<th>Frames (Hz)</th>
<th>Trajectorys</th>
<th>Read User Type</th>
<th>Key Fields</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGSIM(2)</td>
<td>2007</td>
<td>Hou et al., Zheng et al., Qiu et al., DSS et al., Xie et al., Huang et al., Guo et al., Peng et al., Su et al.(3-11)</td>
<td>Road side cameras</td>
<td>the U.S.</td>
<td>Highway, merging section</td>
<td>3</td>
<td>500-640</td>
<td>1.5</td>
<td>5071</td>
<td>5-6</td>
<td>10</td>
<td>9206</td>
<td>car, truck</td>
<td>instantaneous speed and acceleration</td>
<td>The best known, most widely used and earliest published public trajectory dataset</td>
</tr>
<tr>
<td>HighD(12)</td>
<td>2018</td>
<td>Zhang et al., Chen et al., Dong, GU et al.(13-16)</td>
<td>Drone</td>
<td>Germany</td>
<td>Highway, merging section</td>
<td>6</td>
<td>400-420</td>
<td>16.5</td>
<td>45,000</td>
<td>2-3</td>
<td>25</td>
<td>110,000</td>
<td>car, truck</td>
<td>The minimal distance headway(DHW)</td>
<td>The minimal time headway(THW)</td>
</tr>
<tr>
<td>ind(7)</td>
<td>2020</td>
<td>Lu et al., Krasowski et al.(18,19)</td>
<td>Drone</td>
<td>Germany</td>
<td>Unsignalised intersection</td>
<td>4</td>
<td>80<em>40-140</em>50</td>
<td>10</td>
<td>—</td>
<td>2-3</td>
<td>25</td>
<td>15,599</td>
<td>car, truck, pedestrian, bicycle</td>
<td>The type of road users, and their horizontal and longitudinal speed and acceleration</td>
<td>Four different recording locations. Different intersection types. Typical-positioning error &lt;10 cm. HD maps in lanelet2 are provided. Visualization of recorded trajectories</td>
</tr>
<tr>
<td>rund(20)</td>
<td>2020</td>
<td>S. Thal et al., D. Deveaux et al.(21,22)</td>
<td>Drone</td>
<td>Germany</td>
<td>Roundabout</td>
<td>3</td>
<td>140*70</td>
<td>6</td>
<td>—</td>
<td>1-2</td>
<td>25</td>
<td>15746</td>
<td>car, truck, van, pedestrian, bicycle, motorcycle</td>
<td>Accurate visualized trajectories, the type of road users, the direction of every trajectory concerning adjacent time steps, speed and acceleration</td>
<td>Officially released data pre-processing and visualization tools: <a href="https://www.github.com/ika-rwth-aachen/drone-dataset-tools">https://www.github.com/ika-rwth-aachen/drone-dataset-tools</a></td>
</tr>
<tr>
<td>cx(25)</td>
<td>2021</td>
<td>Li X et al., P. Tkachenko et al.(24,25)</td>
<td>Drone</td>
<td>Germany</td>
<td>Highway-entrance and exit</td>
<td>7</td>
<td>420</td>
<td>16.1</td>
<td>27,274</td>
<td>2-4</td>
<td>25</td>
<td>60172</td>
<td>car, truck, van</td>
<td>Velocity and acceleration in the x-y and the radial-latitudinal direction, the width of current lane, whether to change lanes, the DW, TH, TTC and relative speed on current lane</td>
<td>High traffic volume. Rich merging scenarios. Different speed limit scenarios (no-speed limit, 120km/h and 100km/h)</td>
</tr>
<tr>
<td>INTERACTION(26)</td>
<td>2019</td>
<td>Kuan, B et al., Xiaoxi et al.(27,28)</td>
<td>Drone and road side cameras</td>
<td>China, the U.S, Germany, Bulgaria</td>
<td>Merging, lane-changing Unsignalised intersection Signaled intersection Roundabout</td>
<td>12</td>
<td>—</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>—</td>
<td>—</td>
<td>10</td>
<td>10053: 14407 3757 10679</td>
</tr>
<tr>
<td>Ubiquitous Traffic Eyes(29)</td>
<td>2020</td>
<td>Rongxia et al., Wen et al(30,31)</td>
<td>Drone</td>
<td>China</td>
<td>Highway, merging section</td>
<td>6</td>
<td>140-430</td>
<td>0.8</td>
<td>—</td>
<td>5</td>
<td>30</td>
<td>7808</td>
<td>car, truck</td>
<td>the DW, TH, TTC</td>
<td>Time accuracy 0.1s, position accuracy 0.01m. Achieved 100% vehicle detection after manual correction</td>
</tr>
</tbody>
</table>


DATA-DRIVEN METHODS AND EVALUATION

According to Xie et al. (7), there are two types of data-driven LCD models: traditional machine learning-based LCD models and deep learning-based LCD models. Deep learning-based models include those based on deep belief networks (DBN), convolutional neural networks (CNN), long short-term memory neural networks (LSTM), and deep reinforcement learning (DRL). Traditional machine learning-based models include those based on neural networks, support vector machines, and Bayesian filters. Table 2 summarizes the commonly used modeling methods, inputs and outputs, and evaluation criteria for data-driven LCD models.

<table>
<thead>
<tr>
<th>Types</th>
<th>Machine Learning Methods</th>
<th>Literature</th>
<th>Year</th>
<th>Inputs</th>
<th>Outputs</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCD models based on traditional machine learning</td>
<td>Neural Networks</td>
<td>Hunt et al.(32)</td>
<td>1994</td>
<td>$x(t)$ $v(t)$ $\Delta S(t)$</td>
<td>Target lanes and coordinates</td>
<td>Correct classification rate of 70%</td>
</tr>
<tr>
<td></td>
<td>BP Neural Networks</td>
<td>Zheng et al.(4)</td>
<td>2014</td>
<td>$x \Delta v$</td>
<td>Target lanes</td>
<td>Leftward lane change prediction accuracy of 94.6%</td>
</tr>
<tr>
<td></td>
<td>BP Neural Networks</td>
<td>Chen et al.(14)</td>
<td>2022</td>
<td>$v \Delta S$</td>
<td>Whether to change lane</td>
<td>Overall accuracy of 96.5%</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>DOU(6)</td>
<td>2016</td>
<td>$x, \Delta v, \Delta S$</td>
<td>Whether to change lane</td>
<td>Non-merging section-94%, Merging section-78%</td>
</tr>
<tr>
<td></td>
<td>Bayesian Networks</td>
<td>Qiu et al.(5)</td>
<td>2015</td>
<td>$v \Delta v \Delta S$</td>
<td>Whether to change lane</td>
<td>Lane change recognition accuracy of 88.7%</td>
</tr>
<tr>
<td></td>
<td>Deep Learning</td>
<td>Xie et al.(7)</td>
<td>2019</td>
<td>$v \Delta v \Delta S$</td>
<td>Whether to change lane</td>
<td>Prediction accuracy of up to 99.32%, significantly better than the comparison group of BP neural network-based and rule-based models</td>
</tr>
<tr>
<td></td>
<td>Deep Learning</td>
<td>Zhang et al.(13)</td>
<td>2020</td>
<td>$x,v,a,\Delta S,\Delta T$</td>
<td>Target lanes</td>
<td>Prediction accuracy of 98.66%</td>
</tr>
<tr>
<td></td>
<td>Deep Learning</td>
<td>Huang et al.(8)</td>
<td>2020</td>
<td>$x, \Delta v, \Delta S$</td>
<td>Coordinates of the next time sequence</td>
<td>By introducing lane lateral offsets, the accuracy and generalization capability of the proposed model can be improved by about 10%</td>
</tr>
<tr>
<td></td>
<td>Deep Learning</td>
<td>Guo(9)</td>
<td>2021</td>
<td>$x, \Delta v, \Delta S$</td>
<td>Coordinates of the next time sequence</td>
<td>In the accuracy test, the model was reduced by 31% compared to the GRU comparison group. In the mobility test, the MSE was reduced by 39.7%</td>
</tr>
<tr>
<td></td>
<td>Deep Reinforcement Learning</td>
<td>MIRCHEVSKA et al.(35)</td>
<td>2018</td>
<td>$v \Delta S$</td>
<td>Target lanes</td>
<td>Significant improvement in decision-making performance and traffic capacity compared to the rule-based model</td>
</tr>
<tr>
<td></td>
<td>Deep Reinforcement Learning</td>
<td>Peng et al.(10)</td>
<td>2022</td>
<td>$v, \Delta v, a, \Delta S$</td>
<td>Target lanes</td>
<td>24% increase in driving speed compared to original data</td>
</tr>
<tr>
<td>Integrated LCD models</td>
<td>Rule-based+SVM</td>
<td>Ju et al.(11)</td>
<td>2022</td>
<td>$v, \Delta v, \Delta S$</td>
<td>Target lanes</td>
<td>Prediction accuracy improved by 10.78% after augmentation</td>
</tr>
<tr>
<td></td>
<td>Bayesian Networks+Decision Trees</td>
<td>Hou et al.(3)</td>
<td>2014</td>
<td>$\Delta v \Delta S$</td>
<td>Whether to change lane</td>
<td>Prediction accuracy of 79.3% and 94.3% for lane-changing and no lane-changing</td>
</tr>
<tr>
<td></td>
<td>Bayesian Network+BP Neural Networks</td>
<td>Li et al.(36)</td>
<td>2015</td>
<td>the distance to lane lines steering angle</td>
<td>Whether to change lane</td>
<td>Prediction accuracy of 91.4%, improved 6% compared to a merely BP NN based model</td>
</tr>
<tr>
<td></td>
<td>Imitation Learning+XGBoost</td>
<td>Song et al.(37)</td>
<td>2021</td>
<td>$v, \Delta v, \Delta S$</td>
<td>Adjacent lanes' passable status</td>
<td>Significant improvements in safety, traffic efficiency, comfort and speed of strategy learning compared to reinforcement learning alone</td>
</tr>
</tbody>
</table>

Note: $x$ refers to position, $v$ refers to velocity, $a$ refers to acceleration, $d$ refers to driving distance, $\Delta S$ refers to DHW, $\Delta T$ refers to THW, $\Delta v$ refers to relative velocity, $t$ refers to current time.

Table 2: Commonly used methods for data-driven lane-changing models
FUTURE RESEARCH DIRECTIONS

Through the examination of existing literature and relevant research (38), this paper outlines the future research directions for developing data-driven CAVs lane-changing decision models from the following three aspects.

1. Data. a. The deficiencies of current datasets, such as noise, lack of driver characteristic information, short road segment lengths, and limited application scenarios, can lead to problems such as low accuracy, inability to take driver characteristics into account, a lack of generalizability, and an inability to apply to multiple lane-changing scenarios. As a result, there is a need for micro-driving trajectory datasets that are larger in scope, have a wider range of scenarios, and contain both micro-driving trajectory and driver characteristics.
   b. Datasets of mixed traffic flow environments are needed. Currently, mainstream datasets are obtained from environments where almost all vehicles are manually operated. Further research needs to be done to determine how well manual driving cars' interactions with the environment might mimic CAV behavior.

2. Modeling methods. a. Most existing data-driven lane-changing decision models are still based on machine learning methods. It is yet unknown how to apply novel or recently discovered artificial intelligence methods and adjust to the aforementioned newly released data sources. b. Achieving a balance between lowering model complexity and increasing model prediction accuracy and interpretability. Data-driven lane-changing decision models need to consider the most significant variables affecting lane-changing behavior as well as how to build a minimized model.

3. Verification and testing. It may not be enough to simply compare the precision of a single lane-changing decision-making model or assess its ability to replicate traffic phenomena. More thought needs to go into how to test and verify the model from both microscopic and macroscopic viewpoints, as well as how to establish more thorough evaluation indicators or procedures.

CONCLUSIONS

This paper provided a state-of-the-art review of the data-driven lane-changing decision models for CAVs. Knowledge graphs using the keywords were explored, and then, the existing research was summarized and examined in accordance with its data source, characteristics, and the applied data-driven modeling methods. In conclusion, this study offered three major orientations for future study from the perspective of data, modeling methods, and verification.
It would contribute to the development of CAVs decision-making by providing a more sufficient dataset, more effective decision models, and more accurate validation.

ACKNOWLEDGEMENTS

This research was supported by the National Key Research and Development Program of China (Grant No.2019YFE0123800), the National Natural Science Foundation of China (Grant No. 52102380), China Postdoctoral Science Foundation (Grant No. 2021T140325, No.2018M642257), Fundamental Research Funds for the Central Universities (Grant No.330920021140).

REFERENCES

(6) Y. Dou, F. Yan and D. Feng. “Lane changing prediction at highway lane drops using support vector machine and artificial neural network classifiers” (C). 2016 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), Banff, AB, Canada, 2016, pp. 901-906.


(21) S. Thal et al., “Generic Detection and Search-based Test Case Generation of Urban Scenarios based on Real Driving Data” (C). 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 2022, pp. 694-701.


