Improving Trustworthiness of Self-driving Systems

Fahad Alotaibi(✉)

ECS, University of Southampton, Southampton, UK
faa2n19@soton.ac.uk

1 Introduction

Self-Driving Vehicles (SDVs) are considered to be safety-critical system. They may jeopardize the lives of passengers in the vehicle and people in the street, or damaging public property such as the transportation infrastructure. According to the National Transportation Safety Board report [1] of an Uber self-driving crash, the accident was caused by the internal components of SDVs when the AI module failed to detect a victim. The autonomous system was implemented to give a human driver control of a vehicle on the unmanaged areas; however, the driver was distracted and did not react within the appropriate time.

In order to ensure the safety of SDVs, properties of the system must be demonstrated, especially the interactions between autonomous system and human driver in performing driving tasks. Event-B would help to emphasize the properties of the system and ensure a safe transition among multiple components of that system. Event-B uses a variety of extensive tools to support both theorem proving and model checking. According to a survey of formal verification tools [2], Event-B and its toolset (*Rodin*) provide a useful technique to support the goals of a *correct-by-construction* design. Therefore, Event-B and its extensive tools can be used to address issues related to SDVs.

2 Problems, Aims, and Objectives of the Research

Problems: Developers of SDVs are faced with one of the main challenges, specifically, ***establishing techniques for verifying safety properties***. There are three problems related to SDVs from the safety engineering perspective. The challenge in ensuring safety in the SDV starts from (1) *the complexity of the autonomous system*, (2) *the interactions between autonomous functions and the human driver*, and (3) *the ambiguous safety constraints*. **Firstly**, the complexity of SDVs system is determined by the automation level and its autonomous functions. According to SAE International [3], the levels of autonomy are classified from 0–5. The automation levels 1 to 4 (semi-automation) involve a human driver within the driving tasks, while automation level 5 (full automation) does not engage a human driver within the driving tasks. **Secondly**, based on the level of autonomy, the human driver and autonomous system may work together to

A. Raschke et al. (Eds.): ABZ 2020, LNCS 12071, pp. 405–408, 2020.
https://doi.org/10.1007/978-3-030-48077-6_32

perform driving tasks. Therefore, ensuring a secure transition of vehicle control between a human driver and the autonomous system is an important aspect. **Thirdly**, due to the variety of scenarios and faults that might occur in the SDV system, gathering the safety constraints to ensure the functionality of the autonomous system is a challenging task.

Aim and Objectives: The main aim of this research is ***to improve the trustworthiness of SDVs by proposing a new methodology***. This aim can be split into a set of objectives as follows. **The first objective** is to reconsider and analyse the taxonomy requirements [3] in order to identify the safety requirements for the autonomous functions of SDVs on different levels of automation. **The second objective** focuses on the interaction among different components of an SDV at the system level in order to find the potential relationships among these components, especially when a human driver component and autonomous controller component frequently take control of the vehicle. **The third objective** deals with the input and output of autonomous functions and their relationships in order to identify the safety constraints for the autonomous functions.

3 The Current Development and Related Works

3.1 The General Approaches of Using Formal Methods Within the Autonomous System

Due to the complexity of the autonomous system architecture, there are many approaches that have been developed to verify and validate a system from the low- or high-level perspective. At the high level, the concept of a rational agent is used to focus on the autonomous controller, who is responsible for making decisions, and simplify the complexity of SDV system. A rational agent is a software that can perceive its environment via sensors and can explain its intentions [4]. The set of rules can be defined and formally verified into a rational agent entity [4]. In order to use the concept of the rational agent, the logical requirements (rules) must be defined. Consequently, formal methods such as LTL (*linear temporal logic*) can be used to verify the rational agent software.

There are some work in the literature that address the low-level issues of the SDV. The implementation of an SDV mainly relies on machine learning (ML) and its deep neural network (DNN). Although ML algorithms would perform with high accuracy in the image classification task, the work of ML might be affected by the input perturbations of image and lead to an incorrect classification. The input perturbations can be anything such as a 'shadow' and 'weather' which can affect the functionality of the SDV. This kind of manipulation is known as 'adversarial perturbations' [5]. Therefore, Huang et al. [5] proposed an automated verification framework for proving the adversarial robustness of the DNN. This approach is based on applying constraints through the layers of the DNN in order to prevent any misclassifications that might be caused by the adversarial examples. These constraints bound the regions of inputs for all points that are related to the same classification result. *Satisfiability Modulo*

Theory (SMT) is used to develop this verification framework. However, in order to apply this type of approach, the diameter of each region that belongs to a specific classification result must be known, and also the potential adversarial examples must be identified as well.

3.2 The Approaches of Using Event-B for Autonomous System

Constructing the Event-B Models at the System Level for Ensuring Interactions Between the Human Driver and Autonomous Systems: The inspired details of this approach are obtained from the cookbook [6] for the modelling and refinement of control systems. The guidelines mentioned in the cookbook suggest that the phenomenon of a system can be divided into two categories: 1) *variables that identify between environment and controller*, 2) *variables that represent the interaction between human operators and the environment*. There were two contributions related to this approach for modelling the functionality of the autonomous controller: a cruise control system, and lane departure warning system (LDWS) [7]. These autonomous functions belong to a lower level of automation (Level-1). However, reconsidering the ideas that were mentioned in [6,7] might help in either analysing the features of taxonomy requirements or modelling forward to the next automation level.

Constructing the Event-B Models Based on the Safety Constraints of an Autonomous Function: The SDV must implement fail-safe mechanisms, often known as the 'policing function' [8]. The concept of fail-safe mechanisms focuses on the functional requirements and is part of the system requirements. The policing function can check output values of autonomous functions such as an ML model at runtime. The important step of using a validation technique such as a policing function is to demonstrate the safety constraints for the autonomous functions, and are used either to validate the result of autonomous functions or detect failures in the runtime. According to Hoang et al. [8], the concept of metamorphic relationships that aim to discover an expected relationship between inputs and outputs can be used to identify the safety constraints which can be used to build a validation model.

4 Proposed Approach and Future Work

Finding a novel method to extract and identify either the safety constraints or the validation requirements for an autonomous function would be an important task. In order to achieve that, there are three layers that would be used to simplify the complexity of the SDV system as follows: *the specification of the features layer*, *the decision mechanism layer*, and *the actuation layer*. The aim of specification layer is to specify features that would be considered for making a driving decision by modifying the vehicle control variables at the actuation layer. Due to the driving decision might be made by the human-driver or autonomous controller, the local and global features are introduced. The local features focus

on the autonomous functions and its safety constraints, while the global features consider the entire system and try to hold features that might be used when an autonomous function can not perform a driving task.

The local features can be identified from the in-deep knowledge about the expected output of autonomous function. For example, the centring lane lines function tries to keep the vehicle in the road lanes by identifying the lane boundaries and modifying the vehicle control variables. Therefore, '*left and right lane boundary*' and '*Yaw angel*' would be the local features which be controlled and monitored by the local monitor function in order to validate the work of the autonomous function. The local monitor function involves the constraints and procedures that can be used when an autonomous function cannot work as expected. For example, when an autonomous function cannot detect the lane lines, the local monitor function may need to notify the SDV system to use a global feature.

The global features such as '*Driver monitored feature*' and '*emergency stop*' might be applied to ensure the safety of system and avoid any potential mistakes of autonomous function. According to the taxonomy requirements (SAE) [3], the automation levels 1–4 require a human-driver in the loop of automation system. Therefore, it is necessary to implement a system for measuring the awareness level of a human-driver by installing a camera inside a vehicle and monitoring the eyes of driver. The global monitor function would hold global features for establishing a safe transition to the human-driver when an autonomous controller fails to preform the driving task.

Finally, Event-B models can be constructed based on the local and global monitor functions in order to emphasize the main properties of the SDV system. The next step is to apply the proposed approach to a practical case study. We will extend the work of the LDWS [7] to move forward into the next automation level (Level-2) where *a monitored human driver feature* is required.

References

1. National Transportation Safety Board (NTSB) (2018). Preliminary Report HWY18MH010
2. Armstrong, R.C., Punnoose, R.J., Wong, M.H., Mayo, J.R.: Survey of Existing Tools for Formal Verification (2014). https://doi.org/10.2172/1166644
3. SAE J3016: Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. Revision September 2016, SAE International
4. Fisher, M., Dennis, L., Webster, M.: Verifying autonomous systems. Commun. ACM **56**(9), 84–93 (2013). https://doi.org/10.1145/2494558
5. Huang, X., Kwiatkowska, M., Wang, S., Wu, M.: Safety verification of deep neural networks. Technical report (2016). http://arxiv.org/abs/1610.06940
6. Butler, M.: Modelling Guidelines for Discrete Control Systems, Deploy Deliverable D15, D6.1 Advances in Methods Public Document, Chapter 8 (2009)
7. Yeganefard, S., Butler, M.: Structuring functional requirements of control systems to facilitate refinement-based formalisation. ECEASST **46**, 8–11 (2011). https://eprints.soton.ac.uk/337259/1/695-2096-1-PB.pdf
8. Hoang, T.S., Sato, N., Myosin, T., Butler, M., Nakagawa, Y., Ogawa, H.: Policing functions for machine learning systems (2018)