Proceedings of CECNet 2021
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210427

Approach to Define the Reliability of Safety-Related Machine Learning Based Functions in Highly Automated Driving

Pengyu SI, Ossmane KRINI[1], Nadine MÜLLER and Aymen OUERTANI
Institute of Functional Safety, Cyber Security and Artificial Intelligence, Baden-Wuerttemberg Cooperative State University Loerrach, Germany

Abstract. Current standards cannot cover the safety requirements of machine learning based functions used in highly automated driving. Because of the opacity of neural networks, some self-driving functions cannot be developed following the V-model. These functions require the expansion of the standards. This paper focuses on this gap and defines functional reliability for such functions to help the future standards control the quality of machine learning based functions. As an example, reliability functions for pedestrian detection are built. Since the quality criteria in computer vision do not consider safety, new approaches for expression and evaluation of this reliability are designed.

Keywords. Functional safety, self-driving, HAD, reliability, pedestrian detection

1. Introduction

The development of semiconductor industry and artificial intelligence brings us closer to fully automated driving. The transition from hands-on driver assistance to hands-off automated driving, however, requires a huge change of system safety [1], which may rebuild the development process of safety systems in vehicles and redefine the legal responsibility between manufacturers and users. The utilization of artificial neural networks is seen as an effective approach for some or even all tasks in highly automated driving (HAD) [1,2], but the opacity of neural networks challenges the standard development process and functional safety considerations.

In this work, we consider the applicability of existing standards on machine learning based functions applied in HAD vehicles and propose a new kind of reliability to fill the gap of HAD safety requirement, which the current standards do not cover. We build functions to express this reliability mathematically in pedestrian detection and generalize the approaches that can improve this reliability.

Our paper is organized as follows. Section 2 introduces the insufficiency of current standards when applied to high-level self-driving vehicles. In section 3 we limit the scope of application and define the reliability functions for different purposes. Section 4 discusses the enhancement methods of reliability. We summarize our work in section 5.

[1] Corresponding author, Director of Institute of Functional Safety, Cyber Security and Artificial Intelligence located by Baden-Wuerttemberg Cooperative State University Loerrach, Hangstraße 46-50, D-79539 Lörrach, Germany; E-mail: krini@dhbw-loerrach.de.

2. The Conversion of the Functional Safety

Functional safety is an important part of the overall system safety in E/E/PE (Electrical/Electronic/Programmable Electronic) systems. It aims to protect people from unacceptable risk of physical injury or health damage, which is caused by human errors or system failures. In automobile industry, because of the high-level integration of electronic components and its inseparable relationship with safety, the functional safety must be considered in detail in the concept phase and completely validated in the test phase. Thus, the International Organization for Standardization published the ISO 26262: Road vehicles – Functional safety [3] to guide the automobile manufacturers. ISO 21448: Road vehicles – Safety Of The Intended Functionality [4], which is generally abbreviated as SOTIF, on the other hand, concentrates on how a target function is to be specified, developed, verified and validated so that it can be regarded as sufficiently safe.

Over the last decade, the development of neural networks (NN) and deep learning brings breakthroughs in computer vision and robotics, which deeply impacts the autonomous driving seen today both in academia and industry [2]. In automobile industry, the traditional standards for functional safety are no more suitable to guide the design and verification of NN-based safety functions [1,2]. The ISO 26262 follows the famous V-model, which splits the target function into sub-specifications and develops it under those specifications. Such a process is called white-box, the whole data processing is transparent and understandable. On the contrary, the NNs are generally black-box, the mapping between in- and output is a statistical approximation, which is not humanly understandable. The SOTIF aims to address such insufficiency of ISO 26262, but focuses mainly on driver assistance systems rather than HAD systems [1]. In driver assistance systems, the driver shall safely take over the vehicles when failure occurs. However, HAD system itself shall make sure a hazard-free status when failure occurs. Thus, the SOTIF cannot cover the safety requirements of HAD. The prospect of HAD demands more attention on the normalization of the development of NN-based functions in automobile. One of the core jobs for the normalization is bringing the functional safety requirements into HAD functions. Like the requirement of hardware reliabilities in ISO 26262, the reliabilities of functions in HAD system should be defined, so that corresponding ASIL (Automotive Safety Integrity Level) and safety methods can be determined.

The in ISO 26262 defined reliability, however, cannot describe the reliability of the HAD functionalities. In HAD systems, multiple sensors will be used for perception of the environment. Based on the environment modeling, the NNs will be performed to understand the surroundings (e.g. pedestrian, obstacle and parking space) and control the vehicle to drive safely. In this case, unlike traditional mechatronic systems in vehicles, one specific sensor takes responsibility for several tasks and multiple sensors work together for one specific task. Therefore, we propose to use the *functional reliability* to describe the quality of the specific HAD function. *Functional reliability* can be validated and calculated in development phase by using a validation dataset. In future standards, a reliability benchmark can be given to guide the automotive manufacturers. An example is given in the following sections to clarify this concept.

Similarly, the concept of software reliability exists in software engineering [5]. It can also not cover the HAD safety requirements, because different influences on safety of different functions are not considered. Imagine the consequence of a not detected person on the road and a not detected parking space. Different from software reliability, the *functional reliability* is function-oriented and services for safety evaluation.

3. The Reliability of Machine Learning Functions

In this chapter we use pedestrian detection as an example to clarify how the functional reliability of machine learning functions in HAD can be evaluated and how it describes the quality of the function regarding safety.

3.1. Function Boundary

In system safety approaches such as STAMP (Systems-Theoretic Accident Model and Processes), the safety constraints or boundaries must be clearly defined [6]. To analyze the intended function reliability, the function boundaries and use cases need to be clarified. Consequently, defined use cases are only part of the intended function [7].

There are two types of strategies in highly automated driving: End to End and perception-planning-action pipeline [2]. End to End means the algorithm maps the sensor data directly to driving commands without any middle steps, while the classical approach manually decomposes the process into the aforementioned pipeline. NVIDIA believes that handcrafted interfaces ultimately limit performance by restricting information flow through the system and researches continuously on End to End solution, they developed PilotNet to verify the strategy [8]. However, safety requirements for End to End learning approaches are more abstract, formulating and validating measurable performance criteria is significantly more difficult [1]. The decomposed process generally follows the perception and localization → high level path planning → behavior arbitration → motion control pipeline. In each part, several functions will be designed. Therefore, the function boundaries can be narrowly defined, so that the law responsibilities can be clarified when products are put on the market. Thus, considering the executability, the first boundary we formulate is: the reliability of machine learning based pedestrian detection is defined for decomposed HAD strategy.

The training technique is another factor that deeply influences the safety boundary. A neural network can be fully trained and then applied. During application the knowledge, abilities and performance of the network will not be changed. This is so-called offline training. Such training is centralized, all deployments have exactly the same performance. The opposite is online training, the neural network continuously learns from the inputs during application, for this, a mechanism to assess inputs is necessary. Obviously, each deployment of such networks evolves differently because of the individual diversity. This approach is often utilized in reinforcement learning. Clearly, decentralized, online training imposes currently great challenges for safety validation [1]. Thus, we define our second boundary: the considered reliability is aimed at centralized, offline trained neural networks.

Under these constraints we define the pedestrian detection function in HAD. First, we formulate the prioritized definition, which is safety-relevant and shall be verified. Hereafter comes the generalized definition.

- Prioritized definition: detection and tracking of all the pedestrians in front of the driving direction of the automated driving vehicles in urban areas (velocity usually under 60 km/h).
- Generalized definition: detection and tracking of all pedestrians in all directions around the vehicles.

Figure 1 shows a scenario of the use case and the corresponding definition zone.

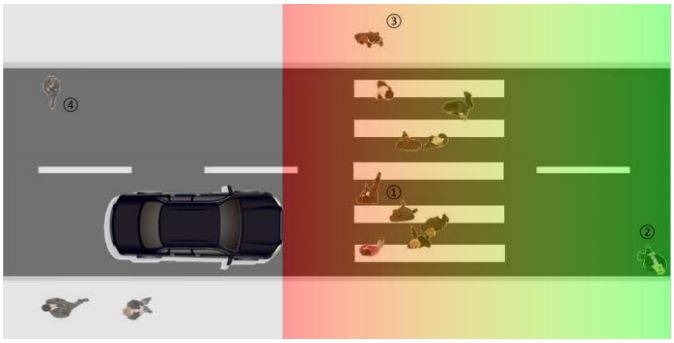

Figure 1. Driving scenario. The red-green colored area satisfies the prioritized definition. The red to green color gradient describes the urgency to take actions. Detection of people in ① demands the highest quality, while prediction ability attenuation in ② is allowed. Interaction with people like ③ belongs to researches in pedestrian behavior understanding [9,10]. Due to its safety irrelevance, detecting ④ is just a task for generalized definition.

3.2. Statistical Reliability of the Function

Camera is the main sensor that is generally used for pedestrian detection. CNN-based neural networks will be used for image processing. Unlike traditional sensor, which directly changes the measured quantity to an electrical signal, the output of NN is a statistical result based on the big training data. Traditional sensors give out a measured value with a confidence interval satisfying gauss distribution, which can be represented with measurement uncertainty. This representation is meaningful in quality management. The prediction from NN is also given confidence, however, this value reflects the goodness of fit of the sample to the fitted function, which cannot match the real probability of a right prediction. Thus, this description is not suitable for reliability.

In computer vision, the following concepts are widely utilized to measure object detection: tp – true positive, fp – false positive, fn – false negative, tn – true negative, Precision (P) – the ratio of tp compared to all detections, Recall (R) – the ratio of tn compared to all ground truths,

$$P = \frac{tp}{tp+fp} \; and \; R = \frac{tp}{tp+fn} \tag{1}$$

and IOU – Intersection Over Union:

$$IOU = \frac{\mathcal{A}(D \cap T)}{\mathcal{A}(D \cup T)} \tag{2}$$

where $\mathcal{A}$ denotes the area of a region in an image, D is the detected bounding box of the object, and T is the area of the bounding box of the matched ground-truth.

Since these concepts quantitatively describe the quality of single-object detection algorithms in different dimensions, we intend to expand them in defining reliability.

The tp and fp or P and R are not mathematically independent. The predefined $IOU_{threshold}$ determines the boundary between tp and fp. A detection is tp, when $IOU > IOU_{threshold}$, and vice versa. Thus, the $IOU_{threshold}$ should be uniformly defined for all pedestrian detection methods in HAD, so that the basic quality of single detection can be defined, and the P and R can be decoupled. Defining $IOU_{threshold}$

requires discussion and verification. A high $IOU_{threshold}$ leads acceptable detection seen as an error, while a low $IOU_{threshold}$ leads to insufficient precision on perception, the reported detection may deviate greatly from the ground truth.

Considering the reaction (e.g. brake, steer, decelerate) time of vehicles, the distance s between objects and the vehicle must be given for evaluating the reliability. The braking distances for the common speed limits of 30km/h and 50km/h in urban areas are 10m and 29m respectively [11]. We suggest using objects between 10 – 20m and 30 – 50m to evaluate the reliability. Other distances like 100m can be used complementarily for uncommon situations.

Precision P influences the passenger comfort of the vehicle. Algorithms with poor precision mark more positive detections than ground truths. The vehicle status will be more frequently adjusted, unnecessary emergency brakes might be operated. Recall R influences pedestrian safety. Undetected ground truths might lead to collisions and accidents. IOU represents the accuracy of detection. The higher the IOU reaches, the more accurately the velocity and distance can be predicted.

Based on the above arguments, we define the algebraic reliability $\mathcal{R}_s$:

$$\mathcal{R}_s = \sum k_i \cdot \mathcal{R}_i \tag{3}$$

where $s = 10m, 30m,$ possibly $100m$

i — P, R and IOU, $\mathcal{R}_P = P$ etc.

k — influence factor, $k_P + k_R + k_{IOU} = 1$

The algebraic reliability describes the functional quality comprehensively. Reliabilities for different distances will be separately represented. Defining k_i is a task similar to defining $IOU_{threshold}$, a balance between comfort and safety must be considered. Regarding the importance of safety, k_R should dominate $\mathcal{R}_s$, followed by k_P.

Furthermore, we define the reliability vector $\boldsymbol{\mathcal{R}}$:

$$\boldsymbol{\mathcal{R}} = \begin{pmatrix} P \\ R \\ IOU \end{pmatrix} \tag{4}$$

The reliability vector $\boldsymbol{\mathcal{R}}$ enables the visualized representation. See Figure 2.

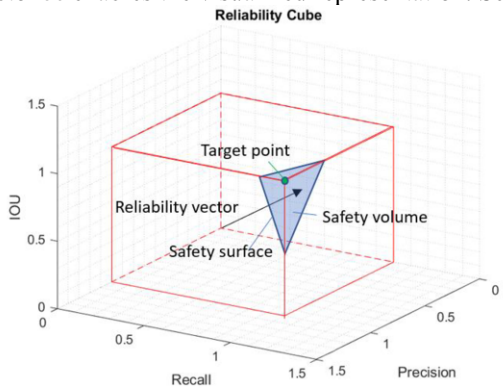

Figure 2. Visualization of reliability vector. A verified acceptable minimum reliability requirement can be shown as safety surface in the figure. The point (1,1,1) is the ideal function. The volume between them is defined as safety volume. Only the pedestrian detection algorithm, whose reliability vector points into this volume, can be seen as functional safe.

Regarding unevenness of above defined k_i, representing the reliability by the magnitude of the vector is not recommended.

Other researchers suggested the utilization of vertical and horizontal pixel deviation between ground truth and detection to judge the confidence level [7]. In this case, the direction influences of bias can be considered, but it complicates the reliability expression.

3.3. Dynamic Reliability of the Function

SSD (Single Shot MultiBox Detector) is nowadays one of the best object detection algorithms, it reaches an average precision (AP) of more than 84.5% for person detection in 512×512 images [12]. The improved algorithm FSSD reaches even 90.2% for the same inputs, with a single NVIDIA 1080Ti GPU, the speed reaches 65.8 FPS for 300×300 input size [13]. Such a score is impressive for normal object detection tasks. Nevertheless, in HAD systems, the missed detections must be handled to satisfy safety requirements. One of the simplest solutions is voting mechanism: Prediction bases on 3 frames, in which the object being in at least two frames as tp marked will be seen as positive. Suppose the AP of FSSD 90.2% is valid for each detection, feed it 3 independent frames, the final AP in this case can be improved significantly to 97.3%.

On the other hand, the pedestrian detection self has limited meaning without the above-mentioned high-level functions like velocity estimation and motion prediction. Such approaches belong to Multi-Object Tracking (MOT) in computer vision. Since HAD is a highly dynamic system, an effective prediction for other surrounded traffic participants can ameliorate both safety and comfort by predictive driving. The standard employed algorithm is tracking-by-detections [14], in which detections in continuous frames are generally basic inputs for tracking functions. The pedestrian detection we defined also includes pedestrian tracking. The deviation of the same object in different frames can influence the prediction quality.

Thus, we propose to validate the dynamic reliability of the function. In MOT there are two widely used metrics to evaluate the tracking quality: CLEAR MOT metrics and ID scores [14]. CLEAR MOT focuses on detection quality, which is similar to the criteria of statistical reliability, while the approach ID scores concentrates on matching cascade. The matching cascade, however, is not a safety-relevant factor in HAD systems. The switch of the IDs of two different pedestrians will not influence the vehicle's motion control. Therefore, both these two metrics are not suitable for dynamic reliability.

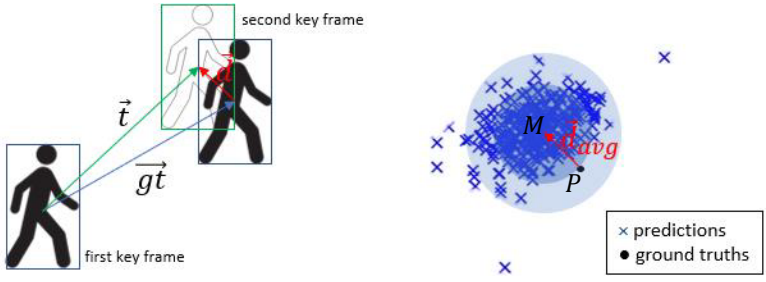

Figure 3. Left: $\overrightarrow{gt}$ is the ground truth motion of the object, $\vec{t}$ is the tracking motion of the object. $\vec{t}$ will be fed to other functions. The key frames are the frames, where the object motion will be estimated. $\vec{d}$ is the deviation vector we defined to describe the tracking quality. Right: Move all the endpoints of $\overrightarrow{gt}$ to P, cross all the endpoints of $\vec{d}$, we see the distribution of all deviations of predictions. The average of $\vec{d}$ is $\vec{d}_{avg}$.

We propose to use the average deviation vector of predictions to describe the prediction accuracy and to use the distribution of the deviations to describe the prediction precision. See Figure 3.

This description can perfectly match the traditional measurement techniques. The $\vec{d}_{avg}$ matches system error and the distance between a cross and M point can be seen as random error. In [15], the pixelwise deviation distribution of YOLOv2 algorithm used in self-driving is investigated. It shows that the x- (horizontal axis) and y-pixel (vertical axis) errors are independently in the x- and y-pixel directions normally distributed, so that the μ and σ are evaluable. Thus, the dynamic reliability can be represented:

$$R_{dym,s} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{5}$$

where $\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$ and s is the above-defined distance.

This definition can benefit the safety functions designed for HAD. Suppose an algorithm has a dynamic reliability μ in driving direction eques 1m and in cross direction eques 0m, σ is 1m, the object is predicted 30m directly in front of the vehicle, it can be concluded that the distance is 95.45% (2- σ) farther than 29m and 99.73% (3- σ) farther than 28m. Using 28m for motion control is 99.73% suitable for this scenario.

This distribution is not proved valid for other algorithms. Regarding the principle of creating bounding box, we assume that YOLO-like algorithms (One-stage) satisfy normal distribution. The distribution of two-stage methods should be investigated. Even if the deviation is not normally distributed, the corresponding probability density function can represent the dynamic reliability and guide the design of safety functions.

4. Improve the Reliability of safety-related Machine Learning Functions

The reliability we proposed aims to reflect the ability of an accurate perception of the driving surroundings. The more accurate the perception is, the safer and more comfortable the vehicles can drive. Thus, the optimization of reliability plays a role in development. Considering the perception process we propose two directions to improve the reliability, namely sensory and algorithmic.

Camera, the main sensor for pedestrian detection, has inherent insufficiencies like low resolution, color sensitivity and dynamic range because of the limited computing power and the performance of light-sensitive elements. Sensor fusion is an effective approach to counteract the inherent insufficiencies of single sensor. A LiDAR, for example, can sense the environment during the camera blindness when driving in and out of the tunnel. Another widely used method to improve functional safety is redundancy. Additional fault tolerance can generally improve a safety integrity level. It is also a practical solution to the blindness of specific camera caused by specific sunlight angles.

Furthermore, the prediction algorithms like LSTM (Long Short-Term Memory) and Kalman Filter can provide a predicted location of the covered object, or when the camera is blinded. With the help of pedestrian behavior understanding, which is based on Bayesian networks, a multi prediction for pedestrian motion can be constructed. The prediction will reference the past location and velocity as well as the intention of the pedestrian. The interaction between vehicles and pedestrians will be built like it in reality.

5. Conclusions and Future Work

In this work, we investigated the evolution of safety requirements with the development of autonomous driving. We expounded on the insufficiency of existing standards applying in HAD and proposed the application of functional reliability. To evaluate and validate the reliability of machine learning based HAD functions, we put forward the definition of a reliability function under predefined constraints to quantitatively describe the reliability of pedestrian detection regarding safety and comfort. At last, we indicated some practical methods, which can improve the defined reliability.

In the future, we will implement and verify some sensor fusion and algorithm methods we mentioned. We are striving to realize a safe, reliable and robust pedestrian detection proposal in HAD. We suggest the automated driving community and automobile industry taking closely part in creating validation datasets and defining key parameters in HAD to accelerate the standardization of HAD vehicles development.

References

[1] Burton S, Gauerhof L, Heinzemann C. Making the Case for Safety of Machine Learning in Highly Automated Driving. In: Tonetta S, Schoitsch E, Bitsch F, editors. Computer Safety, Reliability, and Security: SAFECOMP 2017 Workshops, ASSURE, DECSoS, SASSUR, TELERISE, and TIPS, Trento, Italy, September 12, 2017, Proceedings. Lecture Notes in Computer Science. Vol 10489. Cham: Springer International Publishing; 2017. p. 5–16.

[2] Grigorescu S, Trasnea B, Cocias T, Macesanu G. A Survey of Deep Learning Techniques for Autonomous Driving. J. Field Robotics 2020;37:362–86.

[3] ISO 26262:2011: Road vehicles - Functional safety. Switzerland; 2011.

[4] ISO/PAS 21448:2019: Road vehicles — Safety of the intended functionality. Switzerland; 2019.

[5] Iannino A, Musa JD. Software Reliability. In: Yovits MC, editor. Advances in Computers. New York: Academic Press; 1990. p. 85–170.

[6] Leveson NG. Engineering a safer world: Systems thinking applied to safety. Engineering systems. Cambridge, Massachusetts, London, Englang: The MIT Press; 2017.

[7] Gauerhof L, Munk P, Burton S. Structuring Validation Targets of a Machine Learning Function Applied to Automated Driving. In: Hoshi M, Seki S, editors. Developments in Language Theory: 22nd International Conference, DLT 2018, Tokyo, Japan, September 10-14, 2018, Proceedings. Theoretical Computer Science and General Issues. Vol 11088. Cham: Springer International Publishing; 2018. p. 45–58.

[8] Bojarski M, Chen C, Daw J, Değirmenci A, Deri J, Firner B, et al. The NVIDIA PilotNet Experiments; 2020.

[9] Rasouli A, Kotseruba I, Tsotsos JK. Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. In: 2017 IEEE International Conference on Computer Vision Workshops: ICCVW 2017: 22-29 October 2017, Venice, Italy: proceedings. Piscataway, NJ: IEEE; 2017. p. 206–13.

[10] Rasouli A, Kotseruba I, Tsotsos JK. Understanding Pedestrian Behavior in Complex Traffic Scenes. IEEE Trans. Intell. Veh. 2018;3:61–70.

[11] Greibe P. Braking distance, friction and behaviour Findings, analyses and recommendations based on braking trials 2007.

[12] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. SSD: Single Shot MultiBox Detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision - ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I. Lecture Notes in Computer Science. Vol 9905. Cham, s.l.: Springer International Publishing; 2016. p. 21–37.

[13] Li Z, Zhou F. FSSD: Feature Fusion Single Shot Multibox Detector; 2017.

[14] Ciaparrone G, Luque Sánchez F, Tabik S, Troiano L, Tagliaferri R, Herrera F. Deep learning in video multi-object tracking: A survey. Neurocomputing 2020;381:61–88.

[15] Miethig B, Huangfu Y, Dong J, Tjong J, Mohrenschildt Mv, Habibi S. A Novel Method for Approximating Object Location Error in Bounding Box Detection Algorithms Using A Monocular Camera. IEEE Transactions on Vehicular Technology 2021:1.