Modeling Perception Errors towards Robust Decision Making in Autonomous Vehicles

Andrea Piazzoni[1] , **Jim Cherian**[2] , **Martin Slavik**[2] and **Justin Dauwels**[2,3]
[1]ERI@N, Interdisciplinary Graduate School, Nanyang Technological University, Singapore
[2]Centre of Excellence for Testing & Research of AVs, Nanyang Technological University, Singapore
[3]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
{andrea006, jcherian, martin.slavik, jdauwels}@ntu.edu.sg

Abstract

Sensing and Perception ($S\&P$) is a crucial component of an autonomous system (such as a robot), especially when deployed in highly dynamic environments where it is required to react to unexpected situations. This is particularly true in case of Autonomous Vehicles (AVs) driving on public roads. However, the current evaluation metrics for perception algorithms are typically designed to measure their accuracy per se and do not account for their impact on the decision making subsystem(s). This limitation does not help developers and third party evaluators to answer a critical question: *is the performance of a perception subsystem sufficient for the decision making subsystem to make robust, safe decisions?* In this paper, we propose a simulation-based methodology towards answering this question. At the same time, we show how to analyze the impact of different kinds of sensing and perception errors on the behavior of the autonomous system.

1 Introduction

Autonomous Vehicles (AVs) are, arguably, going to be the first mass deployment of robots that poses a safety impact on public spaces such as roads. The Operational Design Domain (ODD) [SAE, 2018] of an autonomous system may include situations in which some components (or subsystems) exhibit diminished performance, potentially impacting safety. In fact, this is a major safety concern when planning for AV deployments on public roads. For instance, an AV may perform acceptably during daylight hours, but not so when it gets dark. In such situations, we could intuitively infer that the more fallible component is not the decision-making process, but rather the perception subsystem, which may not be able to correctly perceive the surroundings (e.g., without sufficient lighting) thus leading to undesirable AV behavior. On the other hand, we could also infer that the decision making is not robust enough to handle such specific situations [Benenson *et al.*, 2008]. A recent (March 2018) fatal accident of an experimental AV with a jaywalking pedestrian under adverse lighting conditions is a case in point, as the investigation revealed that the decision making was not robust against realistic perception errors [NTSB, 2019]. Evidently, a *mereological* (part-whole) consideration is required, since neither of the subsystems is adequate or inadequate by itself; rather, their combination as a whole is necessary to obtain adequate performance. Therefore, limiting the performance evaluation to individual components does not address the issue of estimating whether the system will be able to operate safely under specific conditions and edge cases. Furthermore, the current metrics designed to evaluate perception are inadequate to answer a critical question: *is the performance of a perception subsystem sufficient for the decision making subsystem to make robust, safe decisions?*

Virtual testing of AVs using simulations offers a safe and convenient way to validate safety [Young *et al.*, 2014]. However, high-fidelity models are necessary to achieve meaningful simulation results that represent the real world. In particular, physics-based sensor simulations may generate synthetic sensor signals to directly challenge the perception. But, they are highly compute-intensive, and therefore, not suitable for the real-time execution of virtual tests under full Automated Driving System (ADS)-in-the-loop or Hardware-in-the-loop configurations. Therefore, it is imperative to develop a feasible alternative that models the intended functionality together with the errors and uncertainty posed by the $S\&P$ subsystem of the ADS, to facilitate virtual testing.

In this paper, we provide some insights towards answering the aforesaid question and make the following contributions:

- We review the state-of-the-art metrics used to measure the performance of AI-based perception algorithms, and identify their limitations in the context of decision making for an autonomous navigation task.
- We recommend some novel directions towards building a representative Perception Error Model (PEM) that can meaningfully describe the performance of the actual sensing and perception of an autonomous system.
- We describe an experimental setup designed to exploit the potential of PEM in a virtual (simulated) environment which offers perfect ground truth, by employing PEMs to replace the actual $S\&P$. By including PEMs, we gain the flexibility to introduce meaningful and representative perception errors while eliminating the need to generate any synthetic sensor signals.
- We demonstrate the usefulness of PEMs as a tool to analyze how the perception capabilities can impact AV behavior, by investigating few representative urban driving

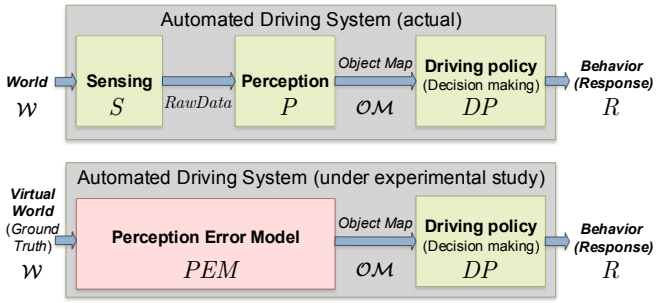


Figure 1: Simplified architecture of an ADS illustrating how PEMs can help to study the impact of $S\&P$ errors on decision making: (top) Actual ADS as it operates in real world (below) ADS with PEM in virtual world.

scenarios. PEMs considered in our experiments also highlight limitations of the standard evaluation metrics.

2 Decision Making Process

Most autonomous systems can be regarded as discrete-time decision making systems that operate in a continuous-time physical world. In the context of AVs, we can term this decision making process as the Driving Policy DP [Shalev-Shwartz *et al.*, 2017] that leads to a physical response, i.e. AV behavior. Although DP can be hand-crafted (based on a rule book), it is tedious and less robust given the complex environment with "surprises" that the AV is expected to operate in. Therefore, many systems learn the art of decision-making from data using reinforcement learning [Shalev-Shwartz *et al.*, 2017], introducing new challenges [Amodei *et al.*, 2016].

Generation of a robust DP for robots operating in controlled environments is a generally tractable problem. This is not so for AVs driving on public roads shared with other traffic participants such as human-driven vehicles and vulnerable road users (e.g., pedestrians or cyclists). Despite the complexity of the ODD, the DP must be sufficiently robust to generate an appropriate real-time AV behavior which is safe as well as comfortable to the occupants. This is arguably a complex dynamic spatio-temporal optimization problem, wherein the constraints possess high *aleatoric* as well as *epistemic* uncertainty [McAllister *et al.*, 2017]. Therefore, the AV research community (including industry, academia and regulators) tries to select scenarios to generate appropriate test cases, and relevant safety metrics that are measurable, objective and robust. Typical metrics include safety clearance distances (between AV and other traffic participants), maximum and minimum limits on AV speed, acceleration, deceleration, jerk and more. Nevertheless, a simple and practical metric is the clearance distance, both in temporal and spatial domains.

3 Error Model

The surrounding environment can be summarized by 3 elements: the map, the ego-vehicle localization, and the other road users or obstacles. In this paper, we focus on the detection of obstacles and other road users. These are described in the Object Map $\mathcal{OM} = \{o_1, \ldots, o_m\}$, the set of the m perceived objects that the $S\&P$ system provides for each frame by observing the surrounding world $\mathcal{W} = \{w_1, \ldots, w_n\}$. The world $\mathcal{W}$ represents the set of n detectable objects (a.k.a. the Ground Truth). The $\mathcal{OM}$ is then analyzed and through the Driving Policy DP, a response R is generated.

To describe each object $w \in \mathcal{W}$, we adopt the same notation as for objects $o \in \mathcal{OM}$ where $o = (\mathbf{X}, \mathbf{C})$:

- **Parameters ($\mathbf{X}$)**: a general vector of parameters such as the (6-9)DoF pose of the object o (3D bounding box),

$$\mathbf{X} = \big(\text{position, rotation, dimensions}\big). \quad (1)$$

  This includes 3 parameters each for position (x, y, z), rotation (yaw, pitch, roll), and dimensions (length, width, height). Some parameters such as pitch, roll, or height may be dismissed in specific road traffic environments. $\mathbf{X}$ can be extended to include any other relevant object parameters such as velocity, turning indicator status, or age (for pedestrians), based on its class $\mathbf{C}$ and depending on the particular ADS and ODD under consideration.

- **Class ($\mathbf{C}$)**: the class of the object, e.g.:

$$\mathbf{C} \in \{Vehicle, Pedestrian\}. \quad (2)$$

  Depending on the system, this can be extended to include other road elements or a finer classification that discriminates between cars, bikes, trucks, etc.

In Figure 1, we note $RawData = S(\mathcal{W})$ and $\mathcal{OM} = P(RawData)$. We can observe that:

$$\mathcal{OM} = P(S(\mathcal{W})) = S\&P(\mathcal{W}) = \mathcal{W} + \mathcal{E}, \quad (3)$$

where $\mathcal{E}$ is the error between $\mathcal{OM}$ and $\mathcal{W}$. The response R is what determines the *behavior* of the AV and therefore, the overall safety and performance of the autonomous system:

$$R = DP(\mathcal{OM}) = DP(\mathcal{W} + \mathcal{E}). \quad (4)$$

The task of assembling the $\mathcal{OM}$ requires to address both classification and regression problems, and has its roots in the object detection task in the Computer Vision (CV) field.

3.1 Evaluation Metrics - State of the Art

The error $\mathcal{E}$ includes predominantly 4 kinds of error. Let o_i be an object perceived corresponding to w_j:

- **False negative**: $o_i \notin \mathcal{OM}$;
- **False positive**: $w_j \notin \mathcal{W}$;
- **Misclassification**: $\mathbf{C}_{o_i} \neq \mathbf{C}_{w_j}$;
- **Parameters errors**: $\mathbf{X}_{o_i} - \mathbf{X}_{w_j} \neq \mathbf{0}$.

All these kinds of errors can be individually observed, statistically measured, and studied by comparing $\mathcal{W}$ and the $\mathcal{OM}$ produced by the $S\&P$. Given Equation 3, the task of analyzing and describing the error is an extension of the task of *measuring* the error of a perception subsystem. In fact, many metrics have been developed for the task of object detection from the CV field. In the field of AVs, many benchmarks on public datasets [Geiger *et al.*, 2012; Huang *et al.*, 2018; Caesar *et al.*, 2019] explore variations of these metrics.

Intersection over Union (IoU) and Mean Average Precision (mAP) are popular metrics for assessing CV algorithms for generic object detection tasks [Everingham *et al.*,

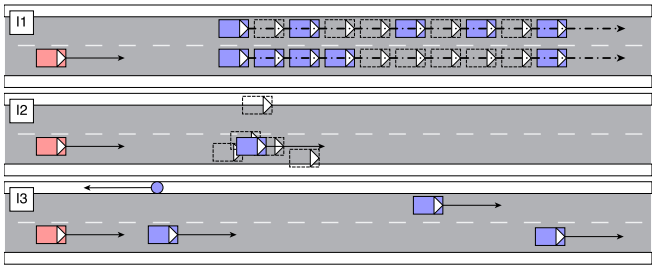


Figure 2: Illustration of the critical issues I1, I2, I3.
I1: Temporal considerations: short vs. long non detection intervals.
I2: Overlap Sensitivity: how sensitive is DP to spatial error?
I3: Relevance of the objects: which ones are active constraints?

2010; Cordts *et al.*, 2016]. Similarly, Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) are common metrics for tracking evaluation [Bernardin and Stiefelhagen, 2008]. All of these metrics require an *a priori* definition of a threshold in order to discriminate between a true positive and a false positive. For example, in [Geiger *et al.*, 2012] the authors consider IoU ≥ 0.7 for the correct detection of a car, or IoU ≥ 0.5 for a pedestrian.

Evaluation Metrics - Critical Issues

In the deployment of AI-based systems, it is not rare that accuracy is not necessarily the best metric to measure their capabilities [Padovani *et al.*, 2019]. While it is not debatable that having a perfect score on accuracy-based metrics is the final goal (i.e., the perception perfectly overlaps with the ground truth, implying $\mathcal{OM} = \mathcal{W}$), these metrics were not designed to consider DP. Hence, they do not provide a model that is adequate enough to study Equation 4. This is because the use of a single metric would hide the specifics of the type of error causing perturbations in the measurement.

In particular, we identify 3 critical areas for analyzing the response R (Figure 2), but are out of the scope of CV metrics:

- **I1: Temporal relevance**: if the system is deployed in a highly dynamic environment, the worst-case error (e.g., losing track of an object for longer intervals) may be more relevant than the average error for same duration.
- **I2: Overlap sensitivity**: The spatial error associated to each object is definitely important. However, considering the bounding box overlap alone may not be sufficient to gauge the quality of the response provided by DP.
- **I3: Relevance of the objects**: Generic CV tasks do not usually associate a weight to each object, as the context may not be considered. However, for an AV in a well-defined ODD, the metrics should judge the relevance of objects considering the context and dynamics (refer I1).

For a more abstract understanding, we should ask: *If the system response R provides the desired outcome, such as avoiding a collision, does it really matter if the $\mathcal{OM}$ had significant errors?* E.g., if the AV brakes to avoid a perceived pedestrian, how much does it matter if the object was actually a cyclist? In this case, how to quantify the relevance of a specific error? There is no straightforward answer to this, since a major classification error could also cause the AV to respond in an unacceptable and/or unsafe manner.

3.2 Error Modeling Considerations

To better understand how the error manifests itself and to subsequently analyze the performance of the $S\&P$, we must first understand the causes of the error.

Positional aspects. Our first observation is that the quality of $S\&P$ is influenced by the **relative position** of w w.r.t. the ego-vehicle, for 2 reasons [Rosique *et al.*, 2019]:

- **Distance**: Performance of all sensors degrades at longer distances. E.g, a more distant object will be captured by fewer pixels by a camera and by fewer LiDAR points.
- **Field of View (FoV)**: Sensors cover different areas around the ego-vehicle. An object that is positioned in an area covered by multiple sensors could be detected with greater accuracy than an object located in an area covered by few, or weaker, sensors.

Parameter inter-dependencies. The second aspect is that the values of any of the object parameters $\mathbf{X}, \mathbf{C}$ can, by themselves, affect the error associated with other parameters of $\mathbf{X}, \mathbf{C}$ [Hoiem *et al.*, 2012]. For example, a larger *size* of an object makes it more likely to be seen at greater distances, whereas the object class $\mathbf{C}$ may limit the error on the size estimation. Some parameters are also not described in the $\mathcal{OM}$ since they are not directly relevant for DP, such as the color or the material of the object [Rosique *et al.*, 2019]. For example, dark/non-reflective surfaces or metallic artifacts may degrade the quality of $\mathcal{OM}$ when S is primarily based on LiDAR or Radar respectively.

Temporal aspects. As a third observation, we can consider the *temporal* aspects of the system, since the DP deals with a sequence of detections, a dynamical system and environment. The overall error of the system can change over time, due to shifting light conditions (e.g., sun blinding, shadows), algorithm uncertainties or even any interference at the level of individual sensors, and should be modeled appropriately. Errors evolve over time and hence should be viewed as time series and be modeled by dynamical models [Mitra *et al.*, 2018].

3.3 Perception Error Model

In this paper, we propose a Perception Error Model (PEM) that represents both the sensing subsystem S and a perception subsystem P, by approximating their combined functions:

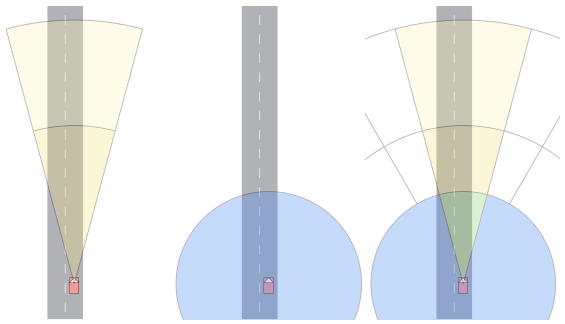
$$PEM(\mathcal{W}) \approx S\&P(\mathcal{W}) = \mathcal{OM} = \mathcal{W} + \mathcal{E}. \quad (5)$$

We propose the following abstraction:

$$PEM = \{\Phi, \Theta, \Gamma\}, \quad (6)$$

where each component is defined as follows:

- Φ: temporal and statistical description of the perception error in function of $\mathbf{X}$;
- Θ: a zone-based spatial description of the $S\&P$ error distribution around the AV considering the coverage by sensors (as illustrated in Figure 3), addressing the positional aspects, viz., the FoV and Distance problems.



(a) Camera (b) LiDAR (c) Overlap

Figure 3: Example of a zone-based partitioning of the $S\&P$ errors. (a) Simple camera FoV, divided into 2 zones based on its range; (b) LiDAR, with 1 zone; (c) Multiple sensors, leading to an overlap region (or regions) where objects can be detected by both sensors. The perception error will depend on how the signals are fused.

- Γ: a description of environmental conditions affecting the error, e.g., a system deployed on the road can be conditioned by the light intensity or fog density, which can be modeled as continuous variables in Φ. Alternatively, one could choose to discretize them and provide distinct Φ for each value (e.g., Φ_{daylight}, Φ_{night}).

Zones-Based Approach for Θ

As illustrated in Figure 3, we propose to address the positional aspects of error by representing the perception error in different *zones*. The FoV problem is easily solved, by dedicating a zone to the overlap of a specific set of sensors. The distance problem instead is already considered in some CV benchmarks [Cordts *et al.*, 2016; Waymo, 2019]. The common solution is to discretize the distance in different ranges, breaking down the metrics into different regions. Our *zones* approach is, in fact, an extension of that approach; while the zones can also be determined by distance thresholds, we also make the entire approach *sensor-agnostic*. Furthermore, we can apply the *zones* approach to study the contextual *relevance* of the objects for a given driving scenario and a planned manoeuvre. Dedicated models for each zone allows to better understand which are the critical areas of the surroundings.

Key Considerations for PEM

If Φ is designed to simply return each object w without any alteration in its parameters, the model is replicating a perfect $S\&P$ system that is able to detect the ground truth. More interestingly, the model can also be designed to not return objects in specific zones $\in \Theta$, thus replicating cases of non-visibility such as blind spots (i.e. the object is not within the range of any sensor, or is occluded [Suchan *et al.*, 2019]).

Considering the above, we propose that designing a PEM is, without loss of generality, a *regression* task, where the goal is to learn the rules and parameters which describe the difference ($\mathcal{E}$) between $\mathcal{W}$ and the $\mathcal{OM}$ generated by the $S\&P$ subsystem (Equation 3), formalized as Φ, Θ, and Γ.

This task is more complex than the typical procedure for computation of evaluation metrics, viz., comparing perception output to the ground truth. As described in Section 3.2, it involves analyzing influence of spatio-temporal dependencies, and object-specific parameter inter-dependencies (covariances) which are relatively under-explored areas. Such aspects are mainly conditioned by the choice and configuration of sensors and perception algorithms. Thus, academic studies focus mostly on a generalized performance evaluation.

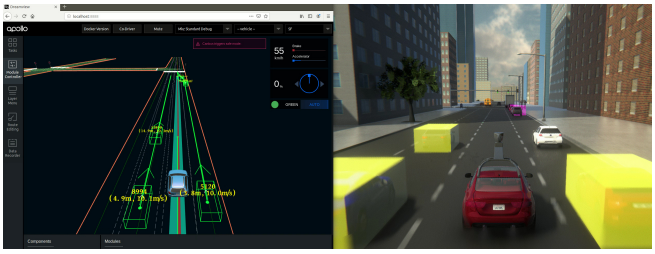


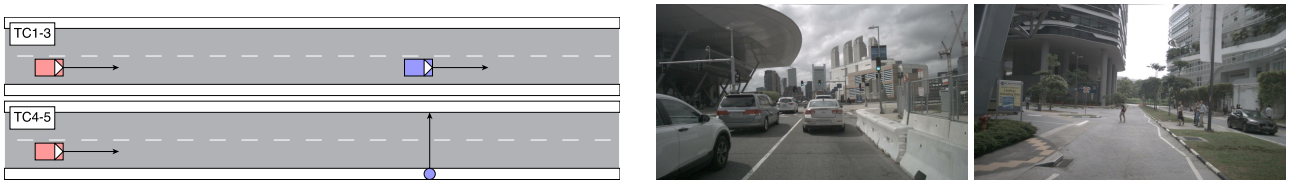
Figure 4: Screenshot of the co-simulation in a generic urban driving situation. Bounding boxes (yellow/purple) of PEM-based objects $\mathcal{OM}$ rendered by LGSVL (right) are consistent with what Apollo sees (left), even undetected objects.

To motivate research in this direction, in the next section we focus on showing how a richer and more descriptive PEM can serve to address some of the issues in the evaluation of both the $S\&P$ and DP. Such an evaluation process is not only crucial from a regulatory perspective, but can also facilitate the system development life-cycle; it can guide developers in choosing the sensors, training the perception models, as well as identifying weaknesses in the DP.

4 Experimental Setup

In this section, we describe the software tools and the experiments conducted to highlight how different kinds of error do (or do not) affect the response R, thus allowing us to observe how a specific PEM can influence R. Furthermore, we describe different PEM variants that serve to demonstrate some of the critical issues discussed in the earlier sections. In this paper, we focus on Φ, i.e., the temporal and statistical description of $S\&P$ error, and specific statistics related to standard evaluation metrics. Since the spatial description Θ and the environmental conditions Γ indirectly facilitate the variations in Φ, we choose not to separately test them in our current experiments. The experiments we designed require a driving simulator and an ADS. To this end, we chose open source tools, namely LGSVL simulator [LGSVL, 2019] and Apollo 3.5 [Fan *et al.*, 2018]. LGSVL simulator is based on the Unity Engine and maintains a reliable bridge between Unity framework and the CyberRT middleware employed by Apollo 3.5, thus enabling co-simulation (see Figure 4). We developed python scripts to implement different scenarios, to automate the tests, configure the simulation environment and the actors in a deterministic manner, and to log the results.

To facilitate our experiments, we adapted these tools so that we could include the PEM in the loop. To this end, we bypassed the built-in $S\&P$ subsystem in Apollo. Firstly, instead of processing (synthetic) raw sensor data, we adapted Apollo to directly read the $\mathcal{OM}$ from a new special-purpose CyberRT topic. Secondly, we defined a new sensor in LGSVL simulator that upon observing $\mathcal{W}$, generates $\mathcal{OM}$ by applying



(a) Illustration of 2 scenarios in our experiments (TC1-3, TC4-5). (b) Following another vehicle. (c) Pedestrian on an urban road.

Figure 5: Scenarios used in experiments: (a) a representative illustration, (b,c) instances from the nuScenes dataset [Caesar *et al.*, 2019].

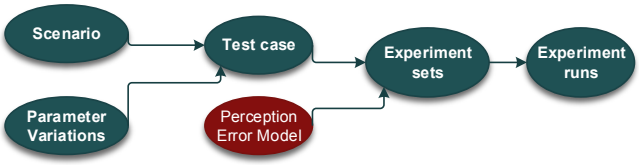


Figure 6: Functional relationship between scenarios, parameter variants, test cases and error models in our experiments.

the specific PEM (see Equation 6) configured for the experiment, and then publishes $\mathcal{OM}$ on the new CyberRT topic that the decision making part of Apollo could read from.

4.1 List of Scenario-based Experiments

In order to study the influence of error models on the AV behavior, we generated a set of experiments following the scheme depicted in Figure 6. In particular, we defined a set of relevant driving scenarios (see Figure 5), configured their parameter variations to obtain concrete test cases, and tested the combinations of each test case with different PEMs to form actual experiment sets. These sets are executed multiple times (at least 30 runs per set), to account for the uncertainty introduced by the randomness involved in our PEMs.

Scenario 1 (test cases TC1-3). It involves an AV driving on a straight road, approaching a traffic vehicle and then following it until they reach a red traffic light. For each test case, each traffic vehicle was set to drive at one of 3 different average speeds, viz. 7, 10, and 15 m/s. To challenge the DP, we applied PEMs that can correctly detect an object ($o_i = w_j$) but randomly fail to include it in $\mathcal{OM}$ for some frames (similar to tracking loss or sporadic non-detections). This allows us to study the critical issue of temporal relevance (see I1).

Implementation TC1-3. We model the False negative errors by means of Markov chains with two states. We tested different values of the parameters *steady state probability* $\in [0.0, 1.0]$ and *mean sojourn time* $\in (0.0s, 10s]$ (average time spent in a state before changing) so as to generate non-detection intervals of varying duration.

Scenario 2 (test cases TC4-5). It is defined by the presence of a pedestrian in 2 different situations: standing in the middle of the road, or jaywalking. For these TCs, we applied PEMs (a) that generate different **positional errors**, with the intention of studying the impact of critical issue I2. We then applied additional PEMs (b) to TC5, so as to replicate the failures that led to a recent AV accident [NTSB, 2019].

Implementation TC4, TC5a. Gaussian White Noise with varying standard deviation σ, applied to the relative position of w w.r.t. the AV, in polar coordinates:

- multiplicative noise on radius d as $\sigma_d \in [0\%, 12\%]$;
- additive noise on azimuth θ as $\sigma_\theta \in [0°, 1.5°]$.

Implementation TC5b. Perfect detection at each frame, but with a tracking loss probability $p_{tl} \in [0, 1]$ for the previously detected obstacles. This may cause the current detection to be considered as a new obstacle, which can hinder the computation of obstacle velocity and lead to unsafe behavior.

5 Experimental Results

In this section, we show the scope of analysis afforded by our experimental setup. Our analysis focuses on the *behavior* of AVs in particular, although the methodology can also be applied elsewhere. For ease of understanding, we show several representative examples from our experimentation. These examples serve to demonstrate the effectiveness of our approach in analyzing how the PEMs can impact the behavior, and thereby taking simple safety metrics under consideration.

5.1 TC1-3: Following a Traffic Vehicle

In Figures 7a, 7b, 7f, 7g, we plot the relationship between two metrics for *behavior evaluation*, namely minimum spatial distance (m) and minimum temporal distance (s), and two statistics of the perception error, namely, relative frequency of detection (realization of the *steady state probability*) and maximum non-detection interval (realization of the *mean sojourn time*). In Figure 7c, we observe that the success rate (no collision) improves with increasing relative detection frequency. Yet, this is true only up to a threshold of $\sim 75\%$, above which the success rate is stationary. On the other hand, in Figure 7h, the duration of the non-detection intervals has a far more significant impact on the success rate. Thus, we may infer that even under low visibility conditions (i.e., low detection probability), if the intervals of non-detection are short enough, the vehicle may be able to avoid collisions. This highlights the importance of including the temporal aspects of the error (critical issue I1) into the evaluation of $S\&P$.

5.2 TC4-5: Pedestrian on an Urban Road

The second scenario offers a different insight. As illustrated in Figures 7d and 7i, we cannot observe major differences in safe behavior (minimum spatial clearance) under varying positional errors generated by different PEMs including the ground truth. This indicates that the system failure is not due to the PEM, but rather due to a weakness in the DP of the

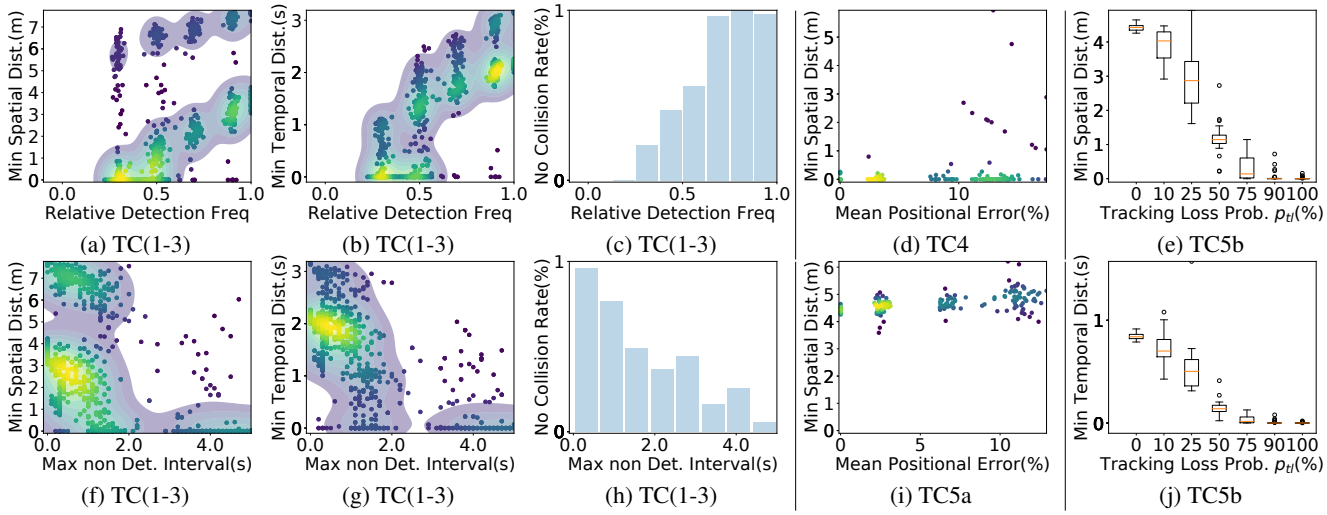


Figure 7: Relationship between safety evaluation metrics and some specific perception statistics. These density scatter plots summarize all the runs (1 dot per run) of the applicable experiment sets, i.e., combinations of a test case and a PEM. Given the high number of samples (up to 900 simulation runs for TC1-3), we highlight the densest areas on a color scale from blue (low density) to yellow (high density).

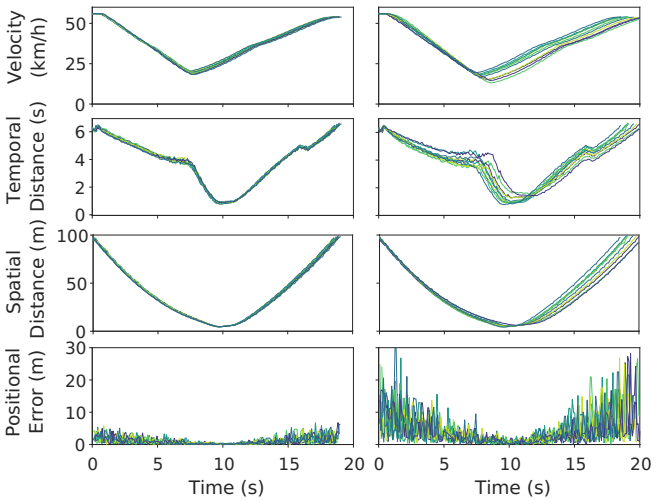


Figure 8: Illustration of ego-vehicle behavior for TC5a with two PEMs: low σ_d, σ_θ (left), and high σ_d, σ_θ (right).

ADS in *our* experimental configuration, which is unable to robustly handle TC4. In fact, safety metrics in TC4 are not influenced by the magnitude of the positional error, since the system fails even with low/no errors. Similarly, in TC5a, the safety is not jeopardized by the error magnitude. This also relates to the other two critical issues I2 and I3 of the current evaluation metrics. Since a positional error by itself can still allow a safe response, it is not adequate to consider IoU as a metric for True Positive. On the contrary, a less restrictive metric should be considered, such as a distance threshold as proposed in [Caesar *et al.*, 2019].

In Figure 8 we compare PEMs with different positional errors in TC5a. Here also, the error magnitude does not have a major impact on safety, although smaller errors can lead to a more consistent AV behavior. Furthermore, experiments for TC5b highlight how the safety decreases as p_{tl} increases, as shown in Figures 7e and 7j. As p_{tl} approaches 0.5 and the obstacle velocity cannot be estimated, DP is unable to predict the obstacle's trajectory and does not brake to avoid it, similar to the AV accident [NTSB, 2019]. This is in contrast to the findings in TC1-3, where frequent tracking errors were easier to handle than infrequent ones. However, it provides an interesting insight towards understanding the contextual relevance of error types depending on the scenario. In particular, in TC1-3 the obstacle is always in the path of the AV, while in cases such as TC5b their paths cross during the scenario (jaywalking in TC5b, but may be similarly applicable for a cut-in scenario). In the latter case, proper obstacle trajectory prediction is critical to foresee the imminent collision.

6 Conclusion

In this paper, we have described a general approach to test and study perception errors in a virtual environment, by linking the respective performance of $S\&P$ and DP, and thereby enabling us to identify their weaknesses. Furthermore, we have implemented an experimental setup to test handcrafted PEMs with the aim of highlighting some limitations of the currently used evaluation metrics for perception algorithms, while discussing how to analyze the resulting system behavior. Although our focus is on AVs, we believe that our approach is general enough to be applied to other domains involving navigational tasks and produce similar insights. The main limitation of the current work lies in the preliminary nature of the experiments. Since no subsystem is adequate or inadequate by itself, the proposed approach has to be adapted to specific AV implementations in order to achieve statistically significant results and safety guarantees. In the near future, we aim to further explore the study of PEMs, develop more realistic simulations that incorporate perception errors, investigate the robustness of $S\&P$ under different environmental conditions, and finally, better approaches to test weaknesses of DP.

References

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[Benenson *et al.*, 2008] Rodrigo Benenson, Thierry Fraichard, and Michel Parent. Achievable safety of driverless ground vehicles. In *2008 10th International Conference on Control, Automation, Robotics and Vision, ICARCV*, pages 515–521. IEEE, dec 2008.

[Bernardin and Stiefelhagen, 2008] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *Eurasip Journal on Image and Video Processing*, 2008, 2008.

[Caesar *et al.*, 2019] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher K I Williams, John Winn, Andrew Zisserman, M Everingham, L KU Van Gool Leuven, Belgium CKI Williams, J Winn, and A Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *Int J Comput Vis*, 88:303–338, 2010.

[Fan *et al.*, 2018] Haoyang Fan, Fan Zhu, Changchun Liu, Liangliang Zhang, Li Zhuang, Dong Li, Weicheng Zhu, Jiangtao Hu, Hongye Li, and Qi Kong. Baidu Apollo EM Motion Planner. *arXiv: 807.08048*, 2018.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[Hoiem *et al.*, 2012] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *Lecture Notes in Computer Science*, number PART 3, pages 340–353. Springer, Berlin, Heidelberg, 2012.

[Huang *et al.*, 2018] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The ApolloScape Open Dataset for Autonomous Driving and its Application. *arXiv: 1803.06184*, 2018.

[LGSVL, 2019] LGSVL. LGSVL Simulator, github.com/lgsvl/simulator, 2019.

[McAllister *et al.*, 2017] Rowan McAllister, Yarin Gal, Alex Kendall, Mark van der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4745–4753, 2017.

[Mitra *et al.*, 2018] Pallavi Mitra, Apratim Choudhury, Vimal Rau Aparow, Giridharan Kulandaivelu, and Justin Dauwels. Towards Modeling of Perception Errors in Autonomous Vehicles. In *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2018.

[NTSB, 2019] NTSB. Collison Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian., 2019. Highway Accident Report NTSB/HAR-19/03. Washington,DC.

[Padovani *et al.*, 2019] Rafael R. Padovani, Lucas N. Ferreira, and Levi H. S. Lelis. Be inaccurate but don't be indecisive: How error distribution can affect user experience. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 2604–2611, 2019.

[Rosique *et al.*, 2019] Francisca Rosique, Pedro J Navarro, Carlos Fernández, and Antonio Padilla. A systematic review of perception system and simulators for autonomous vehicles research. *Sensors*, 19(3):648, 2019.

[SAE, 2018] SAE. J3016_201806 standard: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles, 2018.

[Shalev-Shwartz *et al.*, 2017] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*, 2017.

[Suchan *et al.*, 2019] Jakob Suchan, Mehul Bhatt, and Srikrishna Varadarajan. Out of sight but not out of mind: An answer set programming based online abduction framework for visual sensemaking in autonomous driving. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1879–1885. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[Waymo, 2019] Waymo. Waymo open dataset: An autonomous driving dataset, waymo.com/open, 2019.

[Young *et al.*, 2014] William Young, Amir Sobhani, Michael G Lenné, and Majid Sarvi. Simulation of safety: a review of the state of the art in road safety simulation modelling. *Accident; analysis and prevention*, 66:89–103, 2014.