Symbolic Perception Risk in Autonomous Driving

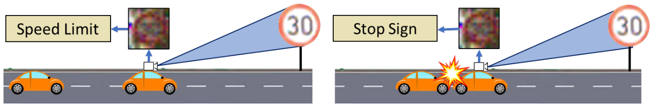
Guangyi Liu, Disha Kamale, Cristian-Ioan Vasile, and Nader Motee

***Abstract*— We develop a novel framework to assess the risk of misperception in a traffic sign classification task in the presence of exogenous noise. We consider the problem in an autonomous driving setting, where detection accuracy gradually improves as the distance to traffic signs decreases due to improved resolution and smaller impact from noise. The common accuracy measures for classification often do not reveal the severity of the potential cost from the misperception. Thus, for the estimated perception statistics obtained using the standard classification algorithms, we aim to quantify the risk of misperception to mitigate the effects of inaccurate detection. By exploring perception outputs, their expected high-level actions, and potential costs, we show the closed-form representation of the conditional value-at-risk (CVaR) of misperception. Moreover, we propose a discounted accumulated CVaR-based risk that leverages the increasing detection quality. Several case studies support the effectiveness of our proposed methodology.**

I. Introduction

"All humans are prone to make mistakes," which are especially crucial while performing safety-critical tasks such as driving. Given the fact that nearly 94% of the accidents are caused by human error [30], and more than 74% among them are related to poor recognition and decision, the development of autonomous driving technologies has gained significant research attention in anticipation of improved human safety [35]. However, due to the inevitable hardware limitations, algorithmic errors as well as external factors such as weather and illumination conditions, it is not rare to see autonomous vehicles performing imperfect recognition of the environment, surrounding vehicles, and making poor decisions that lead to undesirable consequences [8, 3, 10]. For instance, as presented in [11], most autonomous vehicle-related accidents are caused by poor recognition of the environment or surrounding vehicles. Therefore, to maintain the autonomous driving vehicle in a safe operating state in a noisy environment, one must assess the reliability of the noisy belief output.

We consider the motivational scenario when an autonomous vehicle is driving towards a traffic sign, as depicted in Fig. 1. The vehicle is equipped with an onboard camera, which detects, and aims to classify the traffic sign with one of the predefined labels. Due to the location and the unpredictable environment change, the detected image may suffer from various perturbations such as low resolution and pixel-wise noise, which reshape the belief output into a random variable. Our objective is to construct the notion of "risk" of misperception for a given perception model and its visual input, which can be merged into a planning [16] and decision-making module for minimizing the chance of systemic failures [18, 4, 12].



(a) The case when the belief output is *correct* and there is no accident.

(b) The case when *misperception* occurs and its corresponding accident.

Fig. 1: The above diagram depicts the possible outcomes of the misperception.

In the past decades, some well-known risk measures, e.g., Value-at-Risk (VaR) [24] and Conditional Value-at-Risk (CVaR) [25], have shown their significant advances in revealing the uncertainty and reliability of random variables given some harmful tail events. By treating the belief output as a random variable, we use the CVaR measure for the risk quantification, which evaluates the expected outcome when the system has entered the undesired state of operations, e.g., inter-vehicle accidents. The CVaR measure also reveals the severity [33] of the failure when the undesired state is reached, which shows significant importance in our motivational scenario[1]. Consequently, it becomes essential to consider not only the chances of the potential systemic failures but also their magnitudes in terms of costs [15, 21].

The risk quantification process begins with estimating statistics of the belief outputs with a Dirichlet distribution and using the concept of the Voronoi partitioning of the belief space [20] to obtain the discrete distribution of the noisy belief output into class labels. Then, based on the user-defined cost metric for misperceiving each traffic sign, our main result evaluates the risk of traffic sign misperception regarding the severity of the potential accidents.

Our distinct contributions with respect to the existing literature are multi-fold. First, owing to the coherence property, we adopt the Conditional Value-at-Risk (CVaR) measure and propose a risk quantification framework that evaluates the chances of misperception for a given noisy belief output. Secondly, the proposed framework can be customized with a user-defined cost metric and applied to most classification problems when the belief output statistics are available. Thirdly, the proposed approach is control agnostic; given the traffic sign label corresponding to the minimum risk value, our risk-quantification framework is amenable to any control

This work was supported by the ONR N00014-19-1-2478.

G. Liu, D. Kamale, C.I. Vasile and N. Motee are with the Department of Mechanical Engineering and Mechanics, Lehigh University, Bethlehem, PA, 18015, USA. {gliu,ddk320,cvasile,motee}@lehigh.edu.

[1] Mis-perceiving the *Speed-limit of 30 MPH* sign as a *Stop* sign is more dangerous than as a *Speed-limit of 15 MPH* sign.

approach. Finally, the case studies demonstrate a significantly improved performance in terms of safety when using the misperception risk to make decisions under varying levels of noise and resolution.

II. Mathematical Notations

The n-dimensional Euclidean space with elements $\boldsymbol{z} = [z_1, \ldots, z_n]^T$ is denoted by $\mathbb{R}^n$, and $\mathbb{R}_+$ will denote the positive orthant of $\mathbb{R}$. The collection of all integers is denoted by $\mathbb{Z}$, where $\mathbb{Z}_+$ will denote the positive orthant of $\mathbb{Z}$. We represent the $n \times n$ identity matrix as $\boldsymbol{I}_n$ and the vector of all ones as $\mathbf{1}_n$, respectively. The dimension of a vector $\boldsymbol{z} \in \mathbb{R}^n$ is shown by $\dim(\boldsymbol{z}) = n$. The i'th element of a vector $\boldsymbol{x}$ is shown by x_i, and when the vector is time indexed by t, the notation is adapted to $x_{t,i}$. The (i,j)'th entry of matrix $\boldsymbol{A}$ is represented by A_{ij}. Let us define the collection of all feasible belief vectors π, i.e., the belief space, as $\mathcal{P}_n = \{\pi \in \mathbb{R}^n_+ \mid \pi^T \mathbf{1}_n = 1\}$. We also define the Gamma function as $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$ for $\mathbf{Re}(z) > 0$, and the corresponding Digamma function as $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$.

III. Problem Statement

In this paper, we tackle risk evaluation associated with (mis)perception of objects while an autonomous agent is in motion to perform its mission. Objects are semantically linked, and the agent's detection performance depends on its proximity to them.

For the remainder of the paper, let us consider the case of an autonomous vehicle equipped with an onboard camera traveling along a road with traffic signs with labels from $\mathcal{M} = \{1, \ldots, m\}$. The car travels at a constant velocity towards a traffic sign that it will reach in $T \in \mathbb{R}_+$ time. Within the time interval $[0, T]$, the car must take a high-level action from a given set Λ to ensure conformance to traffic rules [6] based on a finite set of observations $\boldsymbol{y}_t$, $t \in [0, T]$, that, ideally, matches the ground truth sign label $\ell \in \mathcal{M}$. The observations $\boldsymbol{y}_t \in \mathbb{R}^{i_d}$ represent images $i_d = q_1 \times q_2$ from the onboard camera in vector form.

We model the detection process as a belief-valued function $h(\cdot)$ over the traffic labels $\mathcal{M}$. Formally, we have $\boldsymbol{p}_t = h(\boldsymbol{y}_t)$, where the *belief output* $\boldsymbol{p}_t \in \mathcal{P}_m$ is the output of the detection algorithm (see §V-A). The *perception output* is the value of function $\arg\max\{\boldsymbol{p}_t\} \in \mathcal{M}$ for a given belief output $\boldsymbol{p}_t \in \mathcal{P}_m$. The functionality of $\arg\max$ can be interpreted as one of the simplest forms of inference in the context of this work.

In most real-world scenarios, the onboard cameras suffer from limited sensing range, resolutions, and noisy visual input. As a result, the generated belief output can be potentially inaccurate and pertaining to noise. The observation model of the image $\boldsymbol{y}_t \in \mathbb{R}^{i_d}$ is given by[2]

$$\boldsymbol{y}_t = g(t, \boldsymbol{y}_0) + b_t \, \boldsymbol{\xi}_t, \tag{1}$$

where $\boldsymbol{y}_0 \in \mathbb{R}^{i_d}$ denotes the high resolution and noise-free image of the traffic sign. These types of perturbation are commonly considered in the research of adversarial attacks of the image classification process using neural network models, see [5, 9, 32].

The nonlinear function $g : [0, T] \times \mathbb{R}^{i_d} \to \mathbb{R}^{i_d}$ modifies the image $\boldsymbol{y}_0$ with various resolutions together with the exogenous disturbances $\boldsymbol{\xi}_t$, which denotes the vector of pixel-wise independent Brownian motions[3] with a time-varying diffusion coefficient b_t. The detail of this modification is further illustrated in §VII-B.

Definition 1. *For a given belief output $\boldsymbol{p}_t \in \mathcal{P}_m$, a misperception occurs if*

$$\arg\max\{\boldsymbol{p}_t\} \neq \ell, \tag{2}$$

where $\ell \in \mathcal{M}$ is the ground truth label of the sign.

The *problem* is to quantify the risk of misperception as a function of the statistics of the noisy belief output $\boldsymbol{p}_t$ and the given confidence level. To reveal the severity of the misperception, we incorporate the user-defined cost metric with the CVaR measure, which connects the event of misperception with the potential loss due to accidents. Once the risk is assessed, the perceived outcome, which corresponds to a traffic sign with the minimal risk level, can be used with any controller of choice, e.g., low-level controller, symbolic controller, or both [19, 16, 28].

IV. Preliminaries

For the exposition of our main result, let us first introduce some necessary results and definitions.

A. The Voronoi Partition of the Belief Space

All possible belief outputs $\boldsymbol{p}_t$ belong to the belief space $\mathcal{P}_m$, which is a $(m-1)-$simplex in $\mathbb{R}^m$. The perception model decides the label of the class is $i \in \mathcal{M}$ if and only if $p_{t,i} > p_{t,j}$ for all $j \in \mathcal{M}$ and $j \neq i$. The $\arg\max$ classification criterion can be explicitly represented via the Voronoi partitioning [17]. The Voronoi partitions of the belief space $\mathcal{P}_m$ are given by $V_1, V_2, \ldots, V_m$, where

$$V_i = \{\boldsymbol{p}_t \in \mathcal{P}_m \mid p_{t,i} > p_{t,j}, \ \forall j \neq i\}. \tag{3}$$

The above partitioning is equivalent to the $\arg\max$ criterion since $\arg\max\{\boldsymbol{p}_t\} = i$ if and only if $\boldsymbol{p}_t \in V_i$, see [20].

B. Conditional Value-at-Risk Measure

To quantify the uncertainty level and the expected outcome encapsulated in belief outputs, we employ the notion of Conditional Value-at-Risk (CVaR) measure [25]. The CVaR indicates the severity of a random variable landing inside an undesirable set of values that characterizes the dangerous state of the system operation with specific confidence level, i.e., Value-at-Risk (VaR). In probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the VaR of the random variable $Y : \Omega \to \mathbb{R}$ is defined as

$$\mathcal{R}^{\varepsilon}_{VaR}(Y) = \min\{z \mid F_Y(z) \geq 1 - \varepsilon\}, \tag{4}$$

where the cumulative distribution function $F_Y(z) = \mathbb{P}\{Y \leq z\}$. Then, the CVaR is defined as follows.

[2] The observation can also be taken from other closed-loop system dynamics, e.g., [7].

[3] One may use other types of noise to model effects of uncertainties.

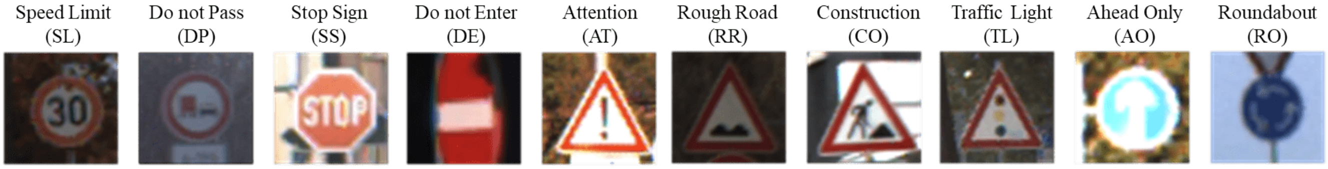


Fig. 2: Ten traffic signs selected from the GTSRB dataset.

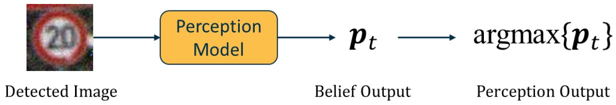


Fig. 3: The traffic sign perception process.

Definition 2. *The Conditional-Value-at-Risk with the confidence level* $(1-\varepsilon) \in [0,1]$ *is the mean of the generalized* $\varepsilon-$*tail distribution:*

$$\mathcal{R}^{\varepsilon}_{CVaR}(Y) = \int_{-\infty}^{\infty} z \; dF^{\varepsilon}_{Y}(z), \tag{5}$$

where

$$F^{\varepsilon}_{Y}(z) = \begin{cases} 0, & \text{if } z < \mathcal{R}^{\varepsilon}_{VaR}(Y) \\ \frac{F_Y(z)+\varepsilon-1}{\varepsilon}, & \text{if } z \geq \mathcal{R}^{\varepsilon}_{VaR}(Y) \end{cases} . \tag{6}$$

A smaller value of ε indicates a higher level of confidence on random variable Y to stay below $\mathcal{R}^{\varepsilon}_{VaR}(Y)$. If Y has a continuous distribution function, CVaR can be obtained as the conditional expectation of Y subject to $Y \geq \mathcal{R}^{\varepsilon}_{VaR}(Y)$. In the case of discrete distributions, one may need to split a probability atom, and CVaR may be obtained by averaging a fractional number of scenarios, see [27].

V. Traffic Sign Perception Model

In this paper, we assume that an off-the-shelf detection algorithm for sign detection is available, e.g., [14, 37, 34], and it can be used to obtain a guess of the sign's ground truth label ℓ at every time $t \in [0,T]$. For simplicity, we assume that input images contain only the signs cropped from the camera's image. Example images of the traffic signs are presented in Fig. 2. We use the German Traffic Sign Recognition Benchmark (GTSRB) dataset[4] to simulate the perception of the vehicle while driving toward the traffic sign. We modify the images from the GTSRB dataset with time-varying resolution and pixel-wise noise, such that the observed image $\boldsymbol{y}_t$ is a random variable. Examples of the modification are presented in §VII-B.

A. Perception Model

The noisy observation $\boldsymbol{y}_t$ is fed into the perception model as an image. In this paper, we consider the perception model as a simple convolutional neural network (CNN) model, e.g., VGG-19 [29], such that $\boldsymbol{p}_t = h(\boldsymbol{y}_t)$ can be rewritten as

$$\boldsymbol{p}_t = \texttt{Softmax}\left(\text{CNN}(\boldsymbol{y}_t)\right), \tag{7}$$

in which belief outputs $\boldsymbol{p}_t = (p_{t,1}, \cdots, p_{t,m}) \in \mathcal{P}_m$ are generated for all continuous time instances within $[0,T]$, as depicted in Fig. 3.

The focus of this work is not to improve the accuracy of detection/recognition algorithms but rather to integrate them with risk quantification to further strengthen the perception module against misperception by reasoning about the severity of the potential failure. For clarity, we choose a simple detection model. However, with simple modifications, the proposed approach applies to any detection model [13, 36].

B. Estimated Statistics of Belief Outputs

To quantify the risk, one must obtain or estimate the statistics of the target random variable. Measuring the statistics of belief output $\boldsymbol{p}_t$ for the entire time-span $[0,T]$ is inefficient due to the time-varying resolution and noise in (1). To resolve this issue, we assume that the statistics of $\boldsymbol{p}_t$ do not change drastically in any sufficiently short time interval $[t-\tau, t) \subset [0,T]$, $\tau \in \mathbb{R}_+$ and $t \in [\tau, T]$, and we can only obtain a finite number of observations in each time interval. Let $\mathcal{T}^{\tau}_t \subset [0,T]$ be the finite set of (uniform or non-uniform) sampling times, and $q = |\mathcal{T}^{\tau}_t|$, the cardinality of the set $\mathcal{T}^{\tau}_t$. Even for the interval $[t-\tau, t)$, quantifying the statistics of $\boldsymbol{p}_t$ from the q observations is not intuitive since the random variable follows the constraint

$$\sum_i p_{t,i} = 1 \text{ and } p_{t,i} \geq 0 \text{ for all } i \in \mathcal{M}, \tag{8}$$

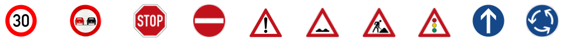
which falls within the belief space $\mathcal{P}_m$. Unlike well-known distribution fitting approaches for Gaussian random variables, we estimate the statistics of $\boldsymbol{p}_t$ using a Dirichlet distribution, for which the corresponding random variable satisfies (8). A random variable $\boldsymbol{z} \in \mathcal{P}_m$ with the Dirichlet distribution $\mathcal{D}(\boldsymbol{z}, \boldsymbol{\alpha})$ has probability density function

$$f_{\mathcal{D}}(z_1, ..., z_m; \alpha_1, ..., \alpha_m) = \frac{\Gamma(\sum_{i=1}^{m} \alpha_i)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \prod_{i=1}^{m} z_i^{\alpha_i - 1}, \tag{9}$$

where $\Gamma(\cdot)$ denotes the Gamma function, $\sum_{i \in \mathcal{M}} z_i = 1$, and $z_i \geq 0$ for all $i \in \mathcal{M}$, and $\boldsymbol{\alpha} \in \mathbb{R}^m_+$ is the concentration parameter vector of the m-order Dirichlet distribution.

Let us consider q images, sampled and processed over a time interval $[t-\tau, t)$, and their corresponding collection of belief outputs is given as a $q \times m$ matrix $\boldsymbol{P}_t = [\boldsymbol{p}^T_{t'}]_{t' \in \mathcal{T}^{\tau}_t}$. We use the fixed point approach proposed in [22] to estimate the Dirichlet distribution, i.e., the concentration vector $\boldsymbol{\alpha}_t$, from a given set of belief outputs $\boldsymbol{P}_t$. The method maximizes the log-likelihood of the estimated distribution and the original data. Considering the convex nature of the problem [26], the

[4] Other datasets, e.g., [23], can also be used with simple modifications.



Sign	SL	DP	SS	DE	AT	RR	CO	TL	AO	RO
SL	0	174	103	103	123	123	121	103	121	120
DP	117	0	105	105	117	117	119	105	97	113
SS	135	109	0	96	110	110	110	96	135	135
DE	117	117	99.5	0	117	500	117	117	117	117
AT	71	111.5	102	92	0	50	0	102	51	137.5
RR	144.5	168	82	82	50	0	50	140	168	258
CO	102	41.5	82	82	30	0	0	41	83	173
TL	97	97	77.5	77.5	39	73	73	0	73	163
AO	91	91	86.5	86.5	45.5	45.5	45.5	91	0	182
RO	83	83	165	165	41.5	41.5	41.5	63	200	0

TABLE I: The cost metric for traffic sign misperception (unit: €1000).

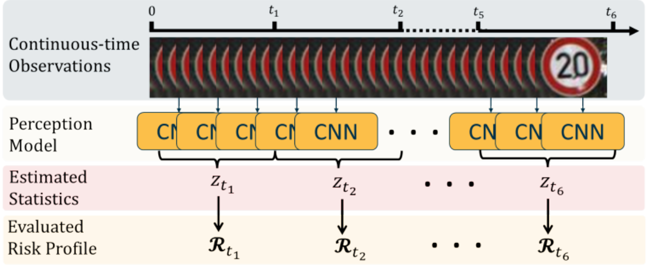


Fig. 4: The above figure illustrates the evaluation of the risk profile for each short time interval, using 6 time intervals as an example.

likelihood is unimodal, and its maximum can be obtained via a simple search [22].

Given an initial guess $\hat{\boldsymbol{\alpha}}_t$ of $\boldsymbol{\alpha}_t$, the estimated value of $\boldsymbol{\alpha}_t$ is updated using the following result.

Lemma 1. *For a given set of belief outputs $\boldsymbol{P}_t$, there exists a set of values of $\boldsymbol{\alpha}_t$ which maximizes the log-likelihood, and $\boldsymbol{\alpha}_t$ can be updated element-wise using*

$$\Psi(\alpha_{t,i}^{new}) = \Psi(\sum_{j=1}^{m} \alpha_{t,j}^{old}) + \frac{1}{q} \sum_{t' \in \mathcal{T}_t^{\tau}} \log p_{t',i}, \tag{10}$$

where $\Psi(\cdot)$ is the Digamma function, $\boldsymbol{\alpha}_t = (\alpha_{t,1}, ..., \alpha_{t,m})$, and $i \in \mathcal{M}$.

The proof of the above result is omitted, and it can be obtained from [22]. The above lemma provides the opportunity of estimating the statistics of belief outputs $\boldsymbol{p}_t$ for the time interval $[t - \tau, t)$ with a Dirichlet random variable $\boldsymbol{z}_t \sim \mathcal{D}(\boldsymbol{z}_t, \boldsymbol{\alpha}_t)$, which allows us to compute the risk of misperception in a closed form.

VI. Risk of Misperception

We introduce the cost for misperceiving traffic signs followed by our main result of misperception risk.

A. Cost of Traffic Signs Misperception

Misperceiving traffic signs often leads to poor decisions of autonomous vehicles, which are primarily associated with high potential costs in real-world driving scenarios. Simply interpreting the belief output as "correct" or "wrong" does not provide adequate information for safe autonomous driving. The reason is that the high-level actions associated with each traffic sign do not yield the same potential cost, e.g., misperceiving the "Speed-limit" sign as a "Stop" sign induces a higher cost than misperceiving it as a "Construction" sign.

Therefore, to explore the severity of the misperception that is beyond correctness, we introduce a cost matrix of misperceiving the traffic signs, $C \in \mathbb{R}^{m \times m}$. For each label $i \in \mathcal{M}$, we define the cost of misperceiving the label $j \in \mathcal{M}$ as the label i as C_{ji}. The correct perception incurs zero cost, i.e., $C_{jj} = 0$. The cost value is user-specific, in this paper, we used the cost metric shown in Table I, obtained by merging the estimated cost for the traffic sign-related violations [2] for our selected traffic signs, such as potential fines and vehicle damages.

B. Risk of Misperceiving Traffic Signs

The cost metric establishes the connection between the estimated belief output $\boldsymbol{z}_t$ to the potential cost of misperception given by the cost metric. For each traffic sign, there exist $m - 1$ possible cases that the estimated perception output $\arg\max\{\boldsymbol{z}_t\}$ is incorrect. For a given estimated belief output $\boldsymbol{z}_t$, one can deterministically define a family of nonlinear functions $r_i(\cdot)|_{i \in \mathcal{M}}$ that maps $\boldsymbol{z}_t$ into the cost metric of misperceptions. For label i, the function $r_i(\cdot)$ takes value in $C_{ji}|_{j \in \mathcal{M}}$ such that

$$r_i(\boldsymbol{z}_t) = C_{ji} \text{ if } \boldsymbol{z}_t \in V_j. \tag{11}$$

Then, the risk quantification can be adjusted to a family of discrete random variables, $r_i(\boldsymbol{z}_t)$ for $i \in \mathcal{M}$.

Given the fact that one or more instances of misperception may obtain the same cost value, let us denote the ordered cost vector as $\boldsymbol{c}_i \in \mathbb{R}^{m'_i}$, where the integer $m'_i \leq m$. The value of m'_i denotes the number of unique values in all $C_{ji}|_{j \in \mathcal{M}}$, and the element of $\boldsymbol{c}_i$ obtains the unique values of C_{ji} in a descending order such that

$$\max_{j \in \mathcal{M}} C_{ji} = (\boldsymbol{c}_i)_1 > \cdots > (\boldsymbol{c}_i)_{m'_i} = \min_{j \in \mathcal{M}} C_{ji}.$$

For the i'th label, there exist m'_i possible cost values for $r_i(\boldsymbol{z}_t)$. Then, the corresponding discrete probability distribution of $r_i(\boldsymbol{z}_t)$ can be computed as follows.

Lemma 2. *For each element of ordered cost vector $\boldsymbol{c}_i$, the probability of $\mathbb{P}\{r_i(\boldsymbol{z}_t) = (\boldsymbol{c}_i)_j\}$ is given by*

$$\mathbb{P}\{r_i(\boldsymbol{z}_t) = (\boldsymbol{c}_i)_j\} = \hat{p}_{t,j} = \sum_{k | C_{k,i} = (\boldsymbol{c}_i)_j} \mathbb{P}\{\boldsymbol{z}_t \in V_k\}, \tag{12}$$

where

$$\mathbb{P}\{\boldsymbol{z}_t \in V_k\} = \int_0^{\infty} \prod_{i \neq k} \left(\frac{\gamma(\alpha_{t,i}, x)}{\Gamma(\alpha_{t,i})} \right) \frac{x^{\alpha_{t,k}-1} \exp(-x)}{\Gamma(\alpha_{t,k})} dx, \tag{13}$$

and $\gamma(\alpha, x)$ is the lower incomplete gamma function.

The integral in (13) can be obtained numerically using the approach proposed in [31] for computing the exceedance probability in the Dirichlet distribution. With the knowledge of the ordered cost vector $\boldsymbol{c}_i|_{i \in \mathcal{M}}$, the estimated statistics of

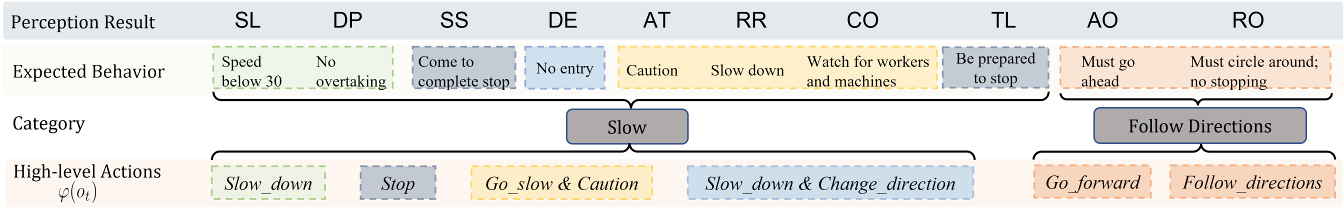


Fig. 5: The high-level actions to be executed upon encountering a specific traffic sign.

belief outputs $\boldsymbol{z}_t$, and its corresponding cost output $r_i(\boldsymbol{z}_t)$, the Conditional Value-at-Risk of traffic sign misperception is shown in the following result.

Theorem 1. *During the time interval $[t-\tau, t)$, given the estimated belief output $\boldsymbol{z}_t$, the risk of misperception with the i'th label is given by*

$$\mathcal{R}^{\varepsilon}_{t,i} = \frac{1}{\varepsilon}\left(\sum_{j=1}^{v}(\boldsymbol{c}_i)_j\,\hat{p}_{t,j} + (\boldsymbol{c}_i)_{v+1}\big(\varepsilon - \sum_{j=1}^{v}\hat{p}_{t,j}\big)\right), \quad (14)$$

where the value of $v \in \mathbb{Z}$ is computed by

$$v = \sup_{v \le m'_i} \sum_{j=1}^{v} \hat{p}_{t,j} \le \varepsilon, \quad (15)$$

the value of $\hat{p}_{t,j}$ is obtained from (12)*, and the value of $1-\varepsilon$ represents the confidence level.*

The above theorem provides the closed-form representation of the risk of misperception when the statistics of the belief output have a Dirichlet distribution. For a given confidence level $1-\varepsilon$, the magnitude of $\mathcal{R}^{\varepsilon}_{t,i}$ quantifies the severity of potential loss when perceiving the traffic sign as the i'th label based on the statistics of $\boldsymbol{z}_t$.

At each discrete time step t, let us also denote the collection of misperception risks for every label as the risk profile of misperception, such that

$$\boldsymbol{\mathcal{R}}^{\varepsilon}_t = [\mathcal{R}^{\varepsilon}_{t,1}, \mathcal{R}^{\varepsilon}_{t,2}, \cdots, \mathcal{R}^{\varepsilon}_{t,m}]^T \in \mathbb{R}^m, \quad (16)$$

and the risk evaluation process is depicted in Fig. 4. The *risk output* is again a label obtained via the $\arg\min$ rule, i.e., $\arg\min\{\boldsymbol{\mathcal{R}}^{\varepsilon}_t\} \in \mathcal{M}$.

C. Accumulated Risk

The risk of misperception is capable of assessing the reliability of the belief output for a time interval $[t-\tau, t)$. However, in the real-world environment, the input quality is time-varying, and simply relying on the risk output for one short time does not reveal the truth about the target traffic sign. Thus, it also requires us to track the change of the risk throughout time and consider both current and past information. To this end, consider a weighted average of risk values with a scaling factor $\mu \in (0,1)$ which balances the importance of the present and the past risk values

$$\hat{\mathcal{R}}^{\varepsilon}_{t,i} = \frac{1-\mu}{1-\mu^K}\sum_{k=1}^{K}\mu^{K-k}\,\mathcal{R}^{\varepsilon}_{k\tau,i}, \quad (17)$$

where $t = K\cdot\tau$ is the current time at step $K \in \mathbb{Z}_+$, $\mathcal{R}^{\varepsilon}_{k\tau,i}$ is evaluated at each time $t' = k\tau$ of step $k \in \{1,\ldots,K\}$ using (14), and $\mu \in (0,1)$ is the user-specified scaling factor. The corresponding risk profile can be stacked as $\hat{\boldsymbol{\mathcal{R}}}^{\varepsilon}_t$ using the same lines of argument in (16).

The accumulated risk shows an evident advantage over the real-time risk $\mathcal{R}_{t,i}$ since it is more inclusive. Moreover, it can handle the situation when the observation changes drastically and the recent observations are unreliable since it does not solely rely on the most recent visual input. The accumulated risk is also normalized to enable tracking of the changes in the risk of misperception, i.e., $\sum_{k=1}^{K}\mu^{K-k} = (1-\mu^K)/(1-\mu)$

Let us also introduce the critical risk threshold $\eta \in \mathbb{R}_+$ as the user-defined maximum acceptable cost of the system. It can be appropriately designed by carefully considering the task and the environmental factors. Once the accumulated risk value $\hat{\mathcal{R}}^{\varepsilon}_{t,i}$ becomes lower than η, the corresponding *accumulated risk output* $o_t \in \mathcal{M}$ is considered as the label, i.e., it is the minimal element in the risk profile such that

$$o_t = \arg\min_{i\in\mathcal{M}}\{\hat{\boldsymbol{\mathcal{R}}}^{\varepsilon}_t\}, \text{ and } \hat{\mathcal{R}}^{\varepsilon}_{t,o_t} \le \eta. \quad (18)$$

VII. Case Study

In this case study, the perception model is trained with the original GTSRB dataset, and the risk of misperception is evaluated with the modified image. Let us consider $\tau = \frac{T}{6}$ and split $[0,T]$ in to 6 intervals for all case studies. We now define some notions used for evaluating the proposed risk metric and then proceed with a detailed discussion on simulations.

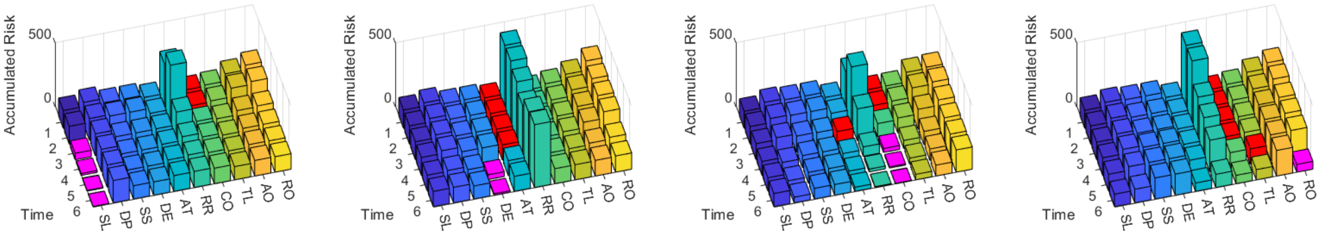
A. High-Level Actions and Time To Execution

We consider a set Λ of prescribed high-level actions that allow the system to execute an appropriate maneuver corresponding to the detected sign as depicted in Fig. 5. The high-level actions are the abstractions of actuation commands to the system.

Given the accumulated risk output o_t, $\varphi : \mathcal{M} \to \Lambda$ provides an action $\lambda \in \Lambda$ to be executed. In order to facilitate the execution of the chosen high-level actions, our framework maximizes the *time to execution* (t_{exec}) defined as follows:

$$t_{exec} = (T-t)\cdot I(\hat{\boldsymbol{\mathcal{R}}}^{\varepsilon}_t \le \eta), \quad (19)$$

where $I(\cdot)$ is an indicator function and takes the value 1 whenever $\hat{\boldsymbol{\mathcal{R}}}^{\varepsilon}_t$ drops below the risk threshold η and is 0 otherwise. Thus, once the normalized accumulated risk



(a) True label: Speed-limit (SL) $\varphi(SL) =$*Slow_down*

(b) True label: Do not enter (DE) $\varphi(DE) =$*Slow_down & Change_direction*

(c) True label: Construction (CO) $\varphi(CO) =$*Go_slow & Caution*

(d) True label: Roundabout (RO) $\varphi(RO) =$*Follow_directions*

Fig. 6: The accumulated risk profile of misperception for various traffic signs $\hat{\boldsymbol{\mathcal{R}}}_t^{\varepsilon}$. At each time step, the label corresponding to minimum risk is shown in red, and the bar indicating o_t is shown in magenta.

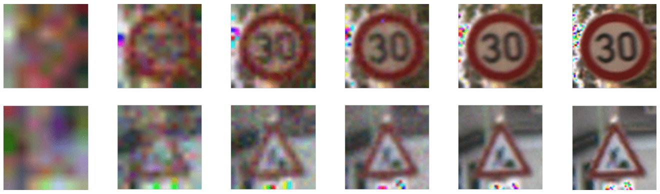
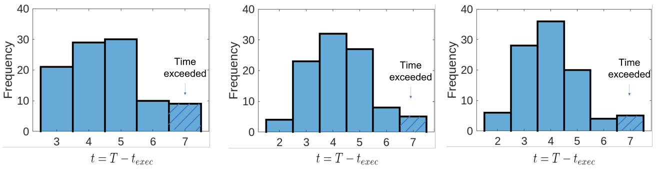


Fig. 7: Example images from the modified GTSRB dataset. Shown for 6 time intervals.



(a) $\eta =$ €1000.

(b) $\eta =$ €10000.

(c) $\eta =$ €50000.

Fig. 8: The distribution of $T - t_{exec}$ with various η. Evaluated among 100 trails.

crosses the acceptable cost η, a corresponding high-level action can be chosen.

As this work focuses on strengthening perception-related safety using risk quantification, we consider a simple mapping between the accumulated risk output o_t to the action space Λ. Developing risk-aware controllers is a topic for future investigation.

B. Modified GTSRB Dataset

Since the images from the GTSRB dataset are static and do not contain any specific types of noise, we apply the following modifications to the dataset, which simulates the scenarios of the vehicle approaching the traffic sign:

- The quality of the detected image is time-varying. The resolution of the detected images increases when the vehicle gets closer to the traffic sign, i.e., as t increases.
- The independent time-varying Gaussian noise is added to each pixel of the image. The magnitude of the noise, b_t, decreases as the vehicle gets closer to the traffic sign, i.e., as t increases.

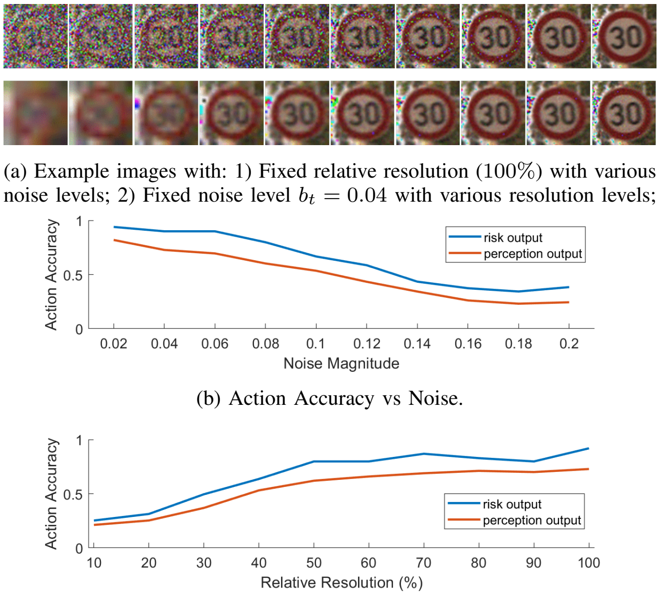
We select 10 types of traffic signs from the dataset, and examples of the modified image are shown in Fig. 7. The above modification can be represented using (1), in which we consider $g(\cdot)$ is the function that changes the resolution w.r.t the time t, and b_t is the time-varying noise magnitude. For instance, our choices are $\dim(g(t, \boldsymbol{y}_0)) = \frac{t}{T}\dim(\boldsymbol{y}_0)$ and $b_t = 0.02\frac{T}{t}$.

C. Simulations

1) Risk of Traffic Sign Misperception: Using the result from Theorem 1, we evaluate the risk of misperceiving specific signs, for which some examples are shown in Fig. 6. The tested images are taken from the modified GTSRB dataset with $\eta =$ €10000 and the confidence level is set by letting $\varepsilon = 0.1$. Among all presented cases, the ground truth labels of the sign $\ell \in \mathcal{M}$ are associated with the accumulated risk output o_t by $t = 6$.

There are a few interesting observations that are worth reporting: The label $\arg\min_{i \in \mathcal{M}}\{\hat{\boldsymbol{\mathcal{R}}}_t^{\varepsilon}\}$ commonly occurs at "CO" when the input image is noisy and with low resolution. This phenomena is because the action related to "CO" is "*Go_slow & Caution*" (see Fig. 5), which is intuitively the safest action when the visual input is not reliable. The severity of different instances of misperceptions depends on the expected high-level actions to be executed, e.g., in Fig. 6c, the high-level actions associated with "AT", "RR", and "CO" are the same, and thus, their corresponding accumulated risks lie within the same range of values. Therefore, the system will not incur a penalty even in case of misperception. In Fig. 6d, the risk output o_t is only obtained at $t = 6$, since the potential loss of misperceiving the label "RO" is relatively high, and the decision should only be made when the visual input is reliable enough, i.e., $\hat{\mathcal{R}}_{t,o_t}^{\varepsilon} \leq \eta$.

2) Critical Risk Threshold: Given different values of the critical risk threshold η, it is expected to observe various distributions of the time to execution t_{exec}, which is presented in Fig. 8. The result indicates that with the increase of the threshold η, the autonomous vehicle could make earlier decisions owing to the higher acceptable potential cost. Thus, the distribution of t_{exec} is more concentrated in higher values, or the distribution of $T - t_{exec}$ is more



(a) Example images with: 1) Fixed relative resolution (100%) with various noise levels; 2) Fixed noise level $b_t = 0.04$ with various resolution levels;

(b) Action Accuracy vs Noise.

(c) Action Accuracy vs Resolution.

Fig. 9: Action accuracy with various noise levels and resolutions.

concentrated toward lower values. Our result also reveals a potential trade-off between η and t_{exec} as a smaller choice of η may prevent autonomous vehicles from making early decisions.

3) Risk Outputs vs Perception Outputs: The risk output $\arg\min_{i\in\mathcal{M}}\{\boldsymbol{\mathcal{R}}_t^{\varepsilon}\}$ exhibits a significant advantage compared to the perception output $\arg\max\{\boldsymbol{p}_t\}$ in the view of preventing the potential losses[5]. We compare the ratio of correct high-level actions, i.e., the action accuracy, generated with both approaches with various noise and resolution levels, as shown in Fig. 9. In both cases, the risk output $\arg\min_{i\in\mathcal{M}}\{\boldsymbol{\mathcal{R}}_t^{\varepsilon}\}$ outperforms the perception output $\arg\max\{\boldsymbol{p}_t\}$ by nearly 20% in the action accuracy. It is because the consequences of each high-level action are inclusively considered in the risk of misperception, which exhibits the major difference between these two approaches.

VIII. Conclusion

In this work, we address the problem of mitigating the effects of misperceiving traffic signs for autonomous driving given noisy visual input. Using the well-known CVaR measure, we construct the framework that evaluates the risk of misperception as a function of the estimated belief output statistics and the user-specified cost metric that captures the severity of potential failures to the system in the event of misperception. Furthermore, leveraging the gradual improvements in the detection accuracy due to gradually improving sensing resolution and smaller effect of noise, we define a discounted accumulated CVaR-based risk. The proposed risk measure reveals the risk output and the accumulated risk output that can be utilized for decision-making and control with any controller of choice. The extensive case studies demonstrate the effectiveness of the proposed risk-quantification framework.

This work is the first step in incorporating the risk of misperception into the perception-planning framework with a focus on the perception module. The immediate natural extension of the work is to design a risk-aware controller to guarantee desirable properties such as system safety and minimum time execution. It also enables the analysis of the perception model from the view of the misperception risk when the output statistics can be explicitly written as a function of model parameters [1].

References

[1] A. Amini, G. Liu, and N. Motee. "Robust Learning of Recurrent Neural Networks in Presence of Exogenous Noise". In: *2021 60th IEEE Conference on Decision and Control*.

[2] M. Ayuso, M. Guillén, and M. Alcañiz. "The impact of traffic violations on the estimated cost of traffic accidents with victims". In: *Accident Analysis & Prevention* 42.2 (2010).

[3] I. Barabás, A. Todoruţ, N. Cordoş, and A. Molea. "Current challenges in autonomous driving". In: *IOP conference series: materials science and engineering*. Vol. 252. 1. IOP Publishing. 2017, p. 012096.

[4] F. S. Barbosa, B. Lacerda, P. Duckworth, J. Tumova, and N. Hawes. "Risk-aware motion planning in partially known environments". In: *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE. 2021, pp. 5220–5226.

[5] F. Carrara, F. Falchi, G. Amato, R. Becarelli, and R. Caldelli. "Detecting Adversarial Inputs by Looking in the black box". In: *arXiv preprint arXiv:1803.02111* (2018).

[6] *Convention on road traffic of 1968 and European Agreement supplementing the convention.* URL: https://unece.org/info/Transport/Road-Traffic-and-Road-Safety/pub/2636.

[7] S. Dean, N. Matni, B. Recht, and V. Ye. "Robust guarantees for perception-based control". In: *Learning for Dynamics and Control*. PMLR. 2020, pp. 350–360.

[8] V. V. Dixit, S. Chand, and D. J. Nair. "Autonomous vehicles: disengagements, accidents and reaction times". In: *PLoS one* 11.12 (2016).

[9] Y. Dong, Q.-A. Fu, X. Yang, T. Pang, H. Su, Z. Xiao, and J. Zhu. "Benchmarking adversarial robustness on image classification". In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 321–331.

[10] M. R. Endsley. "Autonomous driving systems: A preliminary naturalistic study of the Tesla Model S". In: *Journal of Cognitive Engineering and Decision Making* 11.3 (2017).

[11] F. M. Favarò, N. Nader, S. O. Eurich, M. Tripp, and N. Varadaraju. "Examining accident reports involving autonomous vehicles in California". In: *PLoS one* 12.9 (2017).

[12] A. Hakobyan, G. C. Kim, and I. Yang. "Risk-aware motion planning and control using CVaR-constrained optimization". In: *IEEE Robotics and Automation letters* 4.4 (2019).

[13] M. Haloi. "Traffic sign classification using deep inception based convolutional networks". In: *arXiv preprint arXiv:1511.02992* (2015).

[14] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel. "Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark". In: *The 2013 international joint conference on neural networks (IJCNN)*. Ieee. 2013, pp. 1–8.

[15] C. M. Hruschka, D. Töpfer, and S. Zug. "Risk assessment for integral safety in automated driving". In: *2019 2nd International Conference on Intelligent Autonomous Systems*.

[5] The result is validated for a fixed time interval without using accumulated risk to obtain the independent performance for each noise and resolution level.

[16] D. Kamale, S. Haesaert, and C.-I. Vasile. "Cautious Planning with Incremental Symbolic Perception: Designing Verified Reactive Driving Maneuvers". In: *2023 IEEE/RSJ International Conference on Robotics and Automation (ICRA)*.

[17] R. Klein. "Abstract Voronoi diagrams and their applications". In: *Workshop on Computational Geometry*. Springer. 1988.

[18] X. Li, J. DeCastro, C. I. Vasile, S. Karaman, and D. Rus. "Learning A Risk-Aware Trajectory Planner From Demonstrations Using Logic Monitor". In: *Conference on Robot Learning*. PMLR. 2022, pp. 1326–1335.

[19] C. Liu, S. Lee, S. Varnhagen, and H. E. Tseng. "Path planning for autonomous vehicles using model predictive control". In: *2017 IEEE Intelligent Vehicles Symposium (IV)*.

[20] G. Liu, A. Amini, M. Takac, and N. Motee. "Robustness Analysis of Classification Using Recurrent Neural Networks with Perturbed Sequential Input". In: *arXiv preprint arXiv:2203.05403* (2022).

[21] A. Majumdar and M. Pavone. "How should a robot assess risk? towards an axiomatic theory of risk in robotics". In: *Robotics Research*. Springer, 2020, pp. 75–84.

[22] T. Minka. *Estimating a Dirichlet distribution*. 2000.

[23] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder. "The mapillary vistas dataset for semantic understanding of street scenes". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4990–4999.

[24] R. T. Rockafellar and S. Uryasev. "Optimization of Conditional Value-at-Risk". In: *Portfolio The Magazine Of The Fine Arts* 2 (1999), pp. 1–26.

[25] R. T. Rockafellar and S. Uryasev. "Conditional value-at-risk for general loss distributions". In: *Journal of Banking and Finance* 26.7 (2002), pp. 1443–1471.

[26] G. Ronning. "Maximum likelihood estimation of Dirichlet distributions". In: *Journal of statistical computation and simulation* 32.4 (1989), pp. 215–221.

[27] S. Sarykalin, G. Serraino, and S. Uryasev. "Value-at-risk vs. conditional value-at-risk in risk management and optimization". In: *State-of-the-art decision-making tools in the information-intensive age*. Informs, 2008, pp. 270–294.

[28] W. Schwarting, J. Alonso-Mora, and D. Rus. "Planning and decision-making for autonomous vehicles". In: *Annual Review of Control, Robotics, and Autonomous Systems* 1 (2018), pp. 187–210.

[29] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *3rd International Conference on Learning Representations* (2015).

[30] S. Singh. *Critical reasons for crashes investigated in the national motor vehicle crash causation survey*. Tech. rep. 2015.

[31] J. Soch and C. Allefeld. "Exceedance Probabilities for the Dirichlet Distribution". In: *arXiv:1611.01439* (2016).

[32] A. Subramanya, V. Pillai, and H. Pirsiavash. "Fooling network interpretation in image classification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 2020–2029.

[33] C. Wei, M. Fauß, and M. P. Chapman. "CVaR-based Safety Analysis in the Infinite Time Horizon Setting". In: *2022 American Control Conference (ACC)*. IEEE. 2022.

[34] Y. Yang, H. Luo, H. Xu, and F. Wu. "Towards real-time traffic sign detection and classification". In: *IEEE Transactions on Intelligent transportation systems* 17.7 (2015), pp. 2022–2031.

[35] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda. "A survey of autonomous driving: Common practices and emerging technologies". In: *IEEE access* 8 (2020), pp. 58443–58469.

[36] J. Zhang, W. Wang, C. Lu, J. Wang, and A. K. Sangaiah. "Lightweight deep network for traffic sign classification". In: *Annals of Telecommunications* 75.7 (2020), pp. 369–379.

[37] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu. "Traffic-sign detection and classification in the wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2110–2118.