Published as SAE Technical Paper (2017) 2017-01-0050, doi:10.4271/2017-01-0050

Bayesian Test Design for Reliability Assessments of Safety-Relevant Environment Sensors Considering Dependent Failures

Mario Berk, Hans-Martin Kroll, Olaf Schubert, Boris Buschardt, Daniel Straub

Abstract

With increasing levels of driving automation, the perception provided by automotive environment sensors becomes highly safety relevant. A correct assessment of the sensors' perception reliability is therefore crucial for ensuring the safety of the automated driving functionalities. There are currently no standardized procedures or guidelines for demonstrating the perception reliability of the sensors. Engineers therefore face the challenge of setting up test procedures and plan test drive efforts. Null Hypothesis Significance Testing has been employed previously to answer this question. In this contribution, we present an alternative method based on Bayesian parameter inference, which is easy to implement and whose interpretation is more intuitive for engineers without a profound statistical education. We show how to account for different environmental conditions with an influence on sensor performance and for statistical dependence among perception errors. Additionally, we study the impact of error dependence among several sensors on the perception reliability of a redundant multi-sensor system. To this end, we simplify the sensor data fusion with a majority voting scheme, which implies that the multi-sensor system's perception fails whenever more than half of the individual sensors commit unacceptable errors. For a redundant multi-sensor system, in which error occurrence is weakly dependent, it can be shown that empirical reliability assessments are feasible. While the presented method does not encompass entirely the full complexity of the problem, it provides an initial systematic estimate of the necessary test drive effort and facilitates the use of sound statistical methods for test effort estimation.

Introduction

With the advent of advanced driver assistance systems (ADAS) and automated driving, machine vision provided by a set of environment perceiving sensors has become an integral part of modern cars [1–8]. With this technological development arises the need to demonstrate the automated systems' safety and reliability before putting them in service. In this context it is important to assess the reliability of the environment sensors in perceiving the vehicles' surroundings because perception errors may have serious consequences.

The reliability of the automated system and of the sensors' perception depends strongly on the context and environment [9–12]. Existing testing and validation frameworks, for instance ISO 26262, are not directly applicable to the safety validation of automated driving and machine vision [12]. This leads to the task of designing tests to validate the systems' safety. One important question for test design is: How much effort is necessary to empirically demonstrate the reliability of the sensor-based perception, i.e. how much real driving is necessary?

In the context of ADAS and machine vision, a statistical framework utilized to derive the empirical test effort is Null Hypothesis Significance Testing (NHST) [12–15], which is based on the frequentist interpretation of probability. However, for many engineers with only basic training in statistics (and many scientists, see [16]), the correct interpretation of NHST is difficult and counterintuitive. There is a wide discussion in the scientific community about the misuse of NHST based on misinterpretation and overconfidence in "statistical significant" results [16–20].
In this contribution we therefore provide a Bayesian test design for empirical reliability assessments of environment perception. Bayesian methods for reliability assessments are widely used and common in many different industries [21–27].

One important advantage of the Bayesian method compared to the NHST is its flexibility, which allows easier application to non-standard problems such as the reliability of a correlated redundant multi-sensor system. With this method, the necessary test drive effort for a given safety target – e.g. on average less than one safety-relevant perception error in 10^8 hours – can be derived. We find that the Bayesian solution to this problem is easier to apply and interpret for engineers without advanced statistical training. The Bayesian test design is here applied to assess the perception reliability of environment sensors such as Lidar [28] and Radar [29]. Additional to the treatment of individual sensors, the perception reliability of a redundant multi-sensor system is in this study quantified with a majority-voting scheme that has been proposed in [15].

Statistical models utilized to estimate test efforts in the domain of ADAS rely on the assumption of independence of the critical events [12–15]. In this contribution, perception error dependence is taken into account. The main contributions of this work are therefore:

1. The introduction of extensions to standard statistical models with which perception error dependence due to environmental effects can be accounted for at the individual sensor level.
2. The application of well-known and intuitive Bayesian methods to the problem of assessing the perception reliability of automotive environment sensors. The intention is to provide a

detailed guide on perception reliability assessments for practicing engineers.

3. Statistical models for a majority-voting scheme that allow to consider error dependence between redundant sensors are presented in the context of estimating the perception reliability of a multi-sensor system.

The paper is organized as follows: First we review the idea of deriving the test drive effort for reliability demonstrations of ADAS and environment perception with NHST. Thereafter, we introduce a Bayesian approach as an alternative solution to the problem. It includes the dependence of subsequent perception errors and includes non-stationary perception error occurrences. Further, we address how to assess the perception reliability and safety of a redundant multi-sensor system, including error dependence among multiple sensors. A synthetic case study is conducted to demonstrate the proposed methods and to study the impact of error dependence on the reliability of environment perception provided by redundant sensors. Finally, a discussion of the results and the method is given and conclusions are provided.

Background: Reliability assessment of automotive environment perception

The aim of a reliability assessment is to demonstrate that the system or item under consideration – here at first the environment perception of an individual sensor – complies with a given target level of safety. In this specific case, the target level of safety can be expressed as an acceptable mean rate of safety-relevant perception error occurrences, denoted with λ_{SL}:

$$\lambda_{SL} = \frac{1}{\bar{t}} \qquad (1)$$

where $\bar{t}$ is defined as the mean time between the occurrence of subsequent safety-relevant perception errors with the sensor under consideration. The definition of safety-relevant perception errors depends strongly on the ADAS or automated driving functionality of interest. Hence, the safety-relevance of an error has to be determined by the analyst. Generally, a safety-relevant perception error can be the non-detection of an object, the false-positive detection of an object, a large deviation of a physical measurement quantity from the ground truth (e.g. object position or object velocity) or a misclassification of the object (e.g. cyclist identified as pedestrian) [30]. With this definition of potential safety-relevant perception errors, an environment sensors' perception reliability refers to the probability of absence of safety-relevant perception errors, i.e. it is the compliment of the probability of safety-relevant perception error occurrence.

The target level of safety in Eq. (1) may be derived from norms. Alternatively, it was argued that reliability criteria may be defined by requiring the automated driving system to outperform human drivers in terms of the mean rate of accident occurrence [14].

Null Hypothesis Significance Testing for sensor reliability assessment

NHST (see [31] for an introduction to statistics including NHST and [16–20] for the interpretation of NHST results) is a statistical tool utilized to test hypotheses and results in either making a decision in favor or against a tested hypothesis. To be able to make this decision, one sets up a so called null hypothesis H_0 which is opposed to the hypothesis of interest. The hypothesis of interest itself is termed alternative hypothesis H_1. The decision in favor of or against H_1 is based on whether the observed data of an experiment or test are unlikely to occur under the null hypothesis H_0.

When assessing the perception reliability of a sensor, a reasonable H_0 is that the sensor's mean safety-relevant perception error rate λ is larger than the desired target level of safety λ_{SL}:

$$H_0\colon \lambda > \lambda_{SL} \qquad (2)$$

The alternative hypothesis is that the sensor complies with the target level of safety:

$$H_1\colon \lambda \leq \lambda_{SL} \qquad (3)$$

To be able to test H_1 with NHST, it is necessary to specify a random variable Z that summarizes the data (i.e. observations) of a test drive. Z is called the test statistic and is a function of the data. A particular observation of Z is denoted with z_{data}. An obvious choice for Z is the number of safety-relevant perception errors in a test drive. If under H_0, a value of Z smaller or equal to the observed z_{data} is unlikely to occur by chance (i.e. with a probability of less than α, where α denotes the significance level), it will be concluded that with a statistical confidence of $(1-\alpha)$ the null hypothesis can be rejected. With this rejection one implicitly makes a decision in favor of H_1:

$$\text{Reject } H_0 \text{ if } \Pr(Z \leq z_{data}|H_0) = p \leq \alpha \qquad (4)$$

p is the observed significance level of the data. The smaller the observed significance level p, the less likely it is for the observed z_{data} or smaller values to occur by chance, given H_0. It is important to understand that the decision to reject or not to reject H_0 based on $p \leq \alpha$ is also conditional on the statistical assumptions underlying Eq. (4) (e.g. statistical model used, independence, data collection methods).

The necessary test drive effort that allows to make a decision in favor of $H_1\colon \lambda \leq \lambda_{SL}$ can then be derived by the following steps:

- Define a test statistic Z (here: the number of observed perception errors in the test drive)
- Derive the sampling distribution of Z
- Specify a significance level α (typically 0.05)
- Fix z_{data} at different values of the sample space (here: z_{data} is the number of observed errors, with possible values $0,1,2, \ldots$)
- For each value of z_{data}, solve the underlying statistical model of Eq. (4) for the test effort (e.g. number of trials, time or kilometers) such that it holds $\Pr(Z \leq z_{data}\ |H_0) = \alpha$.

Following these steps one obtains the minimum test drive effort associated with the acceptable number of perception errors, such that H_0 is rejected (for an analogous application see [14]).

When selecting a test design, i.e. a value of z_{data} and the corresponding test effort, the error of rejecting a true H_0 – termed type 1 error – is made with a probability α or less, conditional on H_0 and under the condition that the assumptions on the test statistics hold. The type 2 error occurs when not rejecting H_0, even though H_1 is the truth. In case of environment sensors, this error will occur if by chance more safety-relevant errors than acceptable are observed in the predefined test drive effort, even though $\lambda \leq \lambda_{SL}$. The probability β of the type 2 error can be quantified when assuming a specific error rate $\lambda \leq \lambda_{SL}$ to be the hypothetical truth. With a fixed test effort, simultaneously minimizing the type 1 and the type 2 errors is not possible. The lower the type 1 error should be, the larger the type 2 error becomes.

A major problem with NHST after rejecting H_0 is a common but flawed interpretation: Often it is concluded, that at least with $(1-\alpha)$ probability H_1 is true, or with $(1-\alpha)$ probability H_0 is not true [16]. This interpretation is wrong [16–20]. A specific successful hypothesis test that rejects $H_0: \lambda > \lambda_{SL}$ does not demonstrate $H_1: \lambda \leq \lambda_{SL}$ with a certain probability nor does the p-value specify the probability of the data occurring by chance [20]. The hypothesis test only allows the statement that, given the data fulfills Eq. (4), it is not a bad decision to reject H_0, because if H_0 was true, then the observed data would be unlikely. A single hypothesis test makes no statement about the probability of neither H_0 nor H_1 being the truth; in the context of NHST these probabilities are either zero or one [16].

The next section discusses the implications of using NHST for reliability assessments and puts the presented misconception into perspective.

Performance evaluation of NHST

From a societal risk perspective an important question one might ask is: What is the probability of releasing a system with $\lambda > \lambda_{SL}$ (i.e. one that does not comply with the target level of safety), when using the NHST method? The answer to this question depends on how many systems tested with NHST fulfill $\lambda \leq \lambda_{SL}$ in the first place, i.e. on the prior probability $\Pr(H_1)$, and can be estimated as [32]:

$$\Pr(H_0|rejection\ of\ H_0) = \frac{[1-\Pr(H_1)]\cdot\alpha}{\Pr(H_1)\cdot(1-\beta)+[1-\Pr(H_1)]\cdot\alpha} \tag{5}$$

where $\Pr(H_0|rejection\ of\ H_0)$ denotes the ratio of released systems not complying with the target level of safety λ_{SL} to total systems released in the long run. Eq. (5) with its dichotomization of the error rate λ into H_0 and H_1 is in fact a simplification of a continuous probabilistic problem. Therefore, Eq. (5) holds approximately if systems not complying with the target level of safety (i.e. the systems for which H_0 holds) have a λ in the unsafe region $\lambda > \lambda_{SL}$ close to λ_{SL}, and if all systems that fulfill the target level of safety have the same type 2 error probability β. For the case of $\beta = 0.5$, the ratio of released systems with $\lambda > \lambda_{SL}$ to the total number of systems released, after Eq. (5), is illustrated in Figure 1 as a function of the prior probability $\Pr(H_1)$.

First, as Figure 1 shows, if no system complies with the target safety level λ_{SL} in the first place, i.e. $\Pr(H_1) = 0$, then all systems released with NHST fail to comply with the target level of safety. This is a trivial result, but is pointed out here considering the possible misinterpretations of the p-value. Second, if for instance 20 % of the systems tested comply with the target level of safety, i.e. $\Pr(H_1) = 0.2$, then roughly 30 % of all systems released are erroneously considered to comply with the predefined safety requirements. This is far from the significance level $\alpha = 0.05$, which demonstrates how the true error of NHST is easily underestimated when α is misinterpreted to be a probabilistic statement about the tested hypotheses.

In reality, $\Pr(H_1)$ is unknown and the true percentage of released systems not complying with the target level of safety cannot be known with certainty. Nevertheless, Figure 1 demonstrates how one should not be overconfident in NHST results that show the system under consideration is reliable with “statistical significance”.

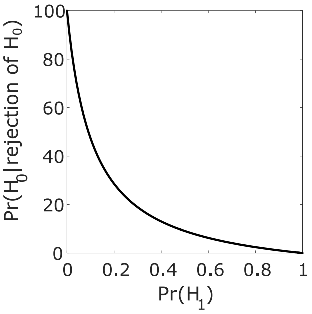

Figure 1. Percentage $\Pr(H_0|rejection\ of\ H_0)$ of released systems with NHST that do not comply with the target level of safety, in function of the prior probability $\Pr(H_1)$ for a system complying with the target level of safety, assuming a type 2 error of $\beta = 0.5$ and a significance level $\alpha = 0.05$.

Alternatives to NHST for reliability assessments

Testing engineers and all involved stakeholders of a reliability analysis might benefit from conceptually easier and more transparent methods than NHST when trying to demonstrate the reliability of new systems such as ADAS or environment sensors. The American Statistical Association (ASA) recently issued a warning about p-values of NHST due to their misuse and misinterpretation [20]: “Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.” Confidence intervals are often put forward as an alternative to NHST. They are related to concepts of NHST and are equally likely to be misinterpreted [16, 33]. Therefore we refrain from further discussing this option.

We find that the Bayesian view on probability is in most contexts better suited for empirical sensor perception reliability evaluation. In contrast to frequentist approaches such as NHST, the Bayesian interpretation of probability treats observed data as fixed and the probabilistic parameters that produced the data as random. The Bayesian approach is often conceptually easier and more directly answers the question usually asked by the analyst: Which probabilistic conclusions about an uncertain parameter of interest can be drawn from a particular set of data or observations [17, 34]? For some problems, a frequentist analysis may yield the same numerical results as a Bayesian analysis [35] but, strictly speaking, does not allow the same intuitive interpretation. The interested reader is referred to [34–36] for further information on the frequentist and Bayesian point of view of probability.

Bayesian methodology for empirical perception reliability assessments of environment sensors

In this section, we propose a Bayesian alternative to NHST to derive the necessary test effort for reliability assessments of environment perception in the field of ADAS. First, a statistical model is presented that accounts for dependent perception errors and for a non-stationary probability of error occurrence. Following the definition of the statistical model, a Bayesian solution to estimate the test drive effort and to perform an empirical sensor perception reliability assessment is derived. Moreover, statistical models are presented that allow to assess the reliability of a multi-sensor system.

Statistical model

Environment sensors such as Lidars [28] and Radars [29] repeatedly probe their environment in measurement cycles and aggregate the collected information in a time-discretized digital environment model containing relevant information about the driving environment and the traffic participants [30]. In this section, the focus is on individual sensors, i.e. the environment representation is not yet based on sensor data fusion [37] but on the data of an individual sensor. Given the temporal discretization of the environmental model, a natural way of describing error occurrence of an individual sensor is to introduce a binary random variable W_i for each measurement cycle i: Either the measurement cycle i is free from safety-relevant perception errors ($w_i = 0$) or at least one safety-relevant perception error occurs ($w_i = 1$). All different perception error types defined previously in the section *Background: Reliability assessment of automotive environment perception* are here considered jointly with the random variable W_i.

With this interpretation, the occurrence of safety-relevant perception errors ($w_i = 1$) in a single measurement cycle i is represented by a Bernoulli trial:

$$p_{W_i}(w_i) = \begin{cases} p & for\ w_i = 1 \\ 1-p & for\ w_i = 0 \end{cases} \tag{6}$$

where p is the probability of error occurrence. The probability of the number of safety-relevant perception errors Y in n measurement cycles can then be modeled with the Binomial distribution:

$$p_Y(y) = \binom{n}{y} \cdot p^y \cdot (1-p)^{n-y} = \cdots$$

$$= \frac{n!}{y! \cdot (n-y)!} \cdot p^y \cdot (1-p)^{n-y} \tag{7}$$

Whenever the probability of error occurrence p is small ($p \to 0$) and the number of measurement cycles is large ($n \to \infty$), both of which holds for environment sensors, in the limit as $n \to \infty$, the Binomial distribution leads to the Poisson distribution:

$$p_{Xt}(x) = \frac{(\lambda \cdot t)^x}{x!} \cdot exp(-\lambda \cdot t) \tag{8}$$

where $x \epsilon [0,1,2, \dots]$ is the number of safety-relevant perception errors in the time interval t and λ is the mean rate of safety-relevant perception error occurrence.

In order for Eqs. (6)–8) to hold, two important requirements have to be met: First, error occurrences in subsequent measurement cycles have to be independent of each other, and second, the probability p and thus the error rate λ have to be constant. Both requirements are not met for environment sensors. The performance of environment sensors such as Lidars or Radars depends on the given context and external factors, including adverse weather conditions, dirt, dust and target properties [9–12]. As a consequence, p and λ are not constant over time. Also, if an error occurs in a given measurement cycle, it will be more likely for the subsequent measurement cycle to exhibit an error due to common influencing factors. Therefore, error occurrence is not independent of each other. Thus, the distributions provided by Eqs. (6)–8) cannot be utilized without violating the underlying mathematical assumptions.

Mathematical representation of dependent errors

The two violations discussed in the previous section are seen to be caused by physical effects that act on different time scales. The perception error dependence is caused by physical effects that are common to multiple measurement cycles in a row. Examples are the presence of objects with low reflectivity, strong rain gusts or a low sun that blinds optical sensors. Due to the highly dynamic nature of driving vehicles, the effects causing the dependence are often only present for a short duration, in the scale of a few seconds. Environment conditions with influence on sensor performance which act on a scale in the order of minutes to hours, such as the weather in general, are not seen as the primary cause of dependent errors but rather influence the overall probability of error occurrence in a specific time interval. These effects consequently lead to a non-stationary error rate and are treated in the next section.

The error dependence leads to a higher probability of error occurrence in subsequent measurement cycles, once an error has occurred. If one wanted to estimate the probability of an perception error occurring for two measurement cycles in a row, with the model given in Eq. (7), neglecting the dependence could lead to severe underestimation. Aside of violating the requirements for Eqs. (6)–8), error dependence is an important factor to consider when assessing the perception reliability of environment sensors because the safety-relevance of errors is partly determined by whether errors persist over multiple cycles (e.g. a false-positive object). A perception error occurring in only one cycle does typically not lead to an insecure or inappropriate behavior of a desired functionality. Sensors are able to use a multi-cycle validation, restricting the impact of errors occurring only for a very short duration [15]. This means that ADAS such as collision protection systems or adaptive cruise control only react when information is consistent over multiple measurement cycles [15, 29, 38].

To account for dependent errors caused by physical effects such as outlined above, the reference of the mean rate of error occurrence λ has to be adapted. In Eq. (8), λ is the rate of safety-relevant errors referring to individual measurement cycles. To consider dependent errors, the error rate λ of Eq. (8) is associated with the interpretation given in Figure 2. F_1, F_2 and F_3 (and so on) are the events that subsequent measurement cycles contain at least one error, at least two errors and at least three errors in a row. Accordingly, λ_1, λ_2 and λ_3 refer to the rate of occurrences of at least one error, at least two errors and at least three errors in a row. Additionally, F_0 denotes the event that a single measurement cycle is free from perception errors. Due to the safety relevance of perception errors that persist over multiple cycles, the analyst ultimately is not interested in λ_1 but rather in λ_2, λ_3 or the rates associated with a larger number of subsequent events.

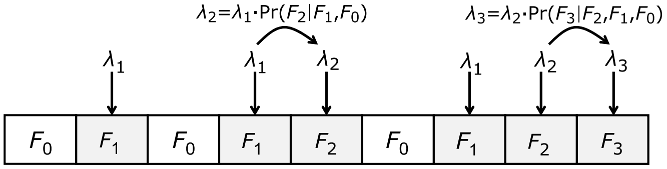

Figure 2. Each box represents a sensor's measurement cycle. A grey colored box indicates that a perception error has occurred in the given cycle. F_1, F_2, F_3 are the events that at least one, at least two and at least three cycles in a row contain an error. λ_1, λ_2, λ_3 are the rates of error occurrences referring to the events F_1, F_2, F_3. F_0 denotes that no error has occurred in a given cycle.

It should be clear, because $\lambda_{\mathrm{j}} \leq \cdots \leq \lambda_2 \leq \lambda_1$, it is easier to learn λ_1 than e.g. λ_2 as more data will be available for λ_1 than λ_2.With the interpretation given in Figure 2, for a fixed number of measurement cycles n, it holds:

$$n = n_{F_0} + n_{F_1} + n_{F_2} + \cdots + n_{F_\infty} \qquad (9)$$

Where n_{F_0} are the number of F_0 events, $n_{F_1}$are the number of F_1 events and so on. When n becomes large $(n \rightarrow \infty)$, n_{F_1} is related to n_{F_0}:

$$n_{F_1} = n_{F_0} \cdot \Pr(F_1|F_0) \qquad (10)$$

where $\Pr(F_1|F_0)$ is the probability that the first cycle of a potential row of errors contains an error, given no error has occurred in the previous cycle (see Figure 2). Similarly, n_{F_j} can be obtained:

$$n_{F_j} = n_{F_0} \cdot \prod_{i=1}^{i=j} \Pr(F_{\mathrm{i}}|F_{\mathrm{i}-1}, \dots, F_0) \qquad (11)$$

$\Pr(F_{\mathrm{i}}|F_{\mathrm{i}-1}, \dots, F_0)$ is the probability of i errors in a row, given $i-1$ erorrs in a row have occurred previously. Thus the dependence of error occurrence is quantified with $\Pr(F_{\mathrm{i}}|F_{\mathrm{i}-1}, \dots, F_0)$ for all $i = 1,2,3, \dots$.Inserting Eqs. (10)(11) into Eq. (9) and solving for n_{F_0}/n leads in the limit of $n \rightarrow \infty$ to the unconditional probability $\Pr(F_0)$ of a randomly selected cycle to be free from perception errors:

$$\Pr(F_0) = \lim_{n \rightarrow \infty} \frac{n_{F_0}}{n} =$$

$$= \frac{1}{1 + \Pr(F_1|F_0) + \Pr(F_1|F_0) \cdot \Pr(F_2|F_1, F_0) + \cdots + \prod_{j=1}^{j=\infty} \Pr(F_{\mathrm{j}}|F_{\mathrm{j}-1}, \dots, F_0)} \qquad (12)$$

Based on $\Pr(F_0)$, λ_1 is obtained as:

$$\lambda_1 = \frac{\Pr(F_0) \cdot \Pr(F_1|F_0)}{t_{cycle}} \qquad (13)$$

With t_{cycle} the measurement cycle time. Generally, it holds:

$$\lambda_{\mathrm{j}} = \lambda_{\mathrm{j}-1} \cdot \Pr(F_{\mathrm{j}}|F_{\mathrm{j}-1}, \dots, F_0) \qquad (14)$$

It follows from Eqs. (12)(14) that λ_{j} fully describes the dependence structure which is quantified by $\Pr(F_0)$, $\Pr(F_1|F_0)$,.., $\Pr(F_{\mathrm{j}}|F_{\mathrm{j}-1}, \dots, F_0)$. Therefore, if the interest is in a sequence of at least j errors in a row, the dependence is fully accounted for. Furthermore, as long as the time intervals are large $(t \rightarrow \infty)$ and the error events are rare, the number of F_{j} events in two non-overlapping time intervals can for a given λ_{j} be regarded as approximately independent of each other. Both these requirements can be assumed to hold for environment sensors. Under these conditions Eq. (8) can be used. Another way of interpreting λ_{j} and $\lambda_{\mathrm{j}-1}$ in Eq. (14) is that they are related by a Poisson process with random selections [39]. In the remainder of the contribution the index j of the λ_{j} of interest will be dropped for ease of notation.

Considering a non-stationary error rate

In this section, the non-stationary rate of error occurrence on a larger time scale is addressed. Environmental conditions and effects with influence on sensor performance that lead to a non-stationary rate of error occurrence are for instance adverse weather, dust, dirt, temperature and many more [9–12].

To account for the non-stationary rate of error occurrence, $\lambda \cdot t$ is in Eq. (8) is replaced by $\mu(t)$:

$$p_{Xt}(x) = \frac{\mu(t)^x}{x!} \cdot exp(-\mu(t)) \qquad (15)$$

$\mu(t)$ is the mean number of safety-relevant errors in the time interval t. For Lidar sensors, weather influences are among the most relevant environmental effects [10, 11]. We therefore use the example of weather conditions to present the calculation of $\mu(t)$. Let the weather be characterized by the four conditions sunny, rainy, snowy and cloudy weather. The sensor performance might differ under different conditions. When the time interval t is large, the mean number of error occurrence $\mu(t)$ can be approximated as:

$$\mu(t) = (p_{sun} \cdot \lambda_{sun} + p_{rain} \cdot \lambda_{rain} + p_{snow} \cdot \lambda_{snow} + \cdots$$

$$+ p_{cloudy} \cdot \lambda_{cloudy}) \cdot t \qquad (16)$$

where $p_{sun}; p_{rain}; p_{snow}$ and p_{cloudy} are the probabilities of sunny, rainy, snowy and cloudy weather. $\lambda_{sun}, \lambda_{rain}, \lambda_{snow}, \lambda_{cloudy}$ are the mean rates of error occurrence during rainy, snowy and cloudy weather. The average error rate $\bar{\lambda}$ can then be calculated as:

$$\bar{\lambda} = p_{sun} \cdot \lambda_{sun} + p_{rain} \cdot \lambda_{rain} + p_{snow} \cdot \lambda_{snow} + \cdots$$

$$+ p_{cloudy} \cdot \lambda_{cloudy} \qquad (17)$$

It has to hold $p_{sun} + p_{rain} + p_{snow} + p_{cloudy} = 1$. Additional environmental effects can be considered by decomposing each error rate $\lambda_{sun}, \lambda_{rain}, \lambda_{snow}, \lambda_{cloudy}$ with respect to the environmental effect that should be added to the model, in analogy to Eq. (17).

To correctly estimate a representative $\bar{\lambda}$, the test drive of total duration t has according to Eqs. (16)(17) be conducted in accordance with $t_{sun} = p_{sun} \cdot t$; $t_{rain} = p_{rain} \cdot t$; $t_{snow} = p_{snow} \cdot t$; $t_{cloudy} = p_{cloudy} \cdot t$. Note that under a varying rate of error occurrence, the error occurrences no longer follows a Poisson process. Nevertheless, the probability of the number of error occurrences can be described by Eq. (8), in which $\lambda \cdot t$ is replaced by $\mu(t)$ (or equivalently by replacing λ with $\bar{\lambda}$).

A problem is however that the probability of for instance p_{rain} varies geographically. For now it is assumed one is able to learn the probabilities $p_{sun}; p_{rain}; p_{snow}$ and p_{cloudy} for one geographical region. Then, for this particular region, the non-stationary rate of error occurrence is accounted for when the test drive is conducted in accordance with $p_{sun}; p_{rain}; p_{snow}$ and p_{cloudy}. A more detailed examination of how to learn $\lambda_{sun}, \lambda_{rain}, \lambda_{snow}, \lambda_{cloudy}$ individually and independent of the geographical region in an efficient way is beyond the scope of this contribution.

Bayesian reliability assessment and test effort estimation

This section describes a Bayesian method for deriving the necessary test drive effort to demonstrate $\lambda < \lambda_{SL}$ before the data is collected and for assessing the reliability of the sensor after the test drive, once the data is available. The general problem is that of inferring an unknown mean rate of safety-relevant perception error occurrence λ from a limited amount of data, where the data consists of the number of safety-relevant perception errors x that have been observed in a specific time interval t. We use Bayesian statistics (see [40] for an introduction) to solve this problem. For a detailed treatment of Bayesian reliability analyses we refer to textbooks [21, 22].

Bayes' theorem is applied to draw probabilistic conclusions on λ:

$$f(\lambda|x,t) \propto f(\lambda) \cdot p_{Xt}(x|\lambda,t) \tag{18}$$

$f(\lambda|x,t)$ is the posterior probability distribution of the safety-relevant perception error rate λ for a given observed number of safety-relevant errors x in the time interval t, $f(\lambda)$ is the prior probability distribution of the error rate λ and $p_{Xt}(x|\lambda,t)$ is the likelihood of λ given the observation of x in t. The likelihood $p_{Xt}(x|\lambda,t)$ is defined by the Poisson distribution of Eq. (8). The symbol ∝ in Eq. (18) expresses that the posterior distribution is proportional to the prior and likelihood up to a constant.

A convenient choice for the prior distribution in case of a Poisson likelihood is the Gamma distribution. The Gamma distribution is the conjugate distribution to the Poisson likelihood, which signifies that both $f(\lambda)$ and $f(\lambda|x,t)$ in Eq. (14) have the Gamma distribution [40]. The Gamma probability density function (PDF) is:

$$f(\lambda) = \frac{b^a}{\Gamma(a)} \cdot \lambda^{a-1} \cdot \exp(-b \cdot \lambda) \tag{19}$$

Where a and b are the parameters of the gamma distribution and $\Gamma(a) = \int_0^\infty u^{a-1} \cdot \exp(-u)\,\mathrm{d}u$ is the gamma function. The corresponding Gamma cumulative distribution function (CDF) $F(\lambda|x,t)$ is:

$$F(\lambda|x,t) = \frac{\gamma(a, b \cdot \lambda)}{\Gamma(a)} \tag{20}$$

Where $\gamma(a, b\cdot\lambda) = \int_0^{b\cdot\lambda} u^{a-1} \cdot \exp(-u)\,\mathrm{d}u$ is the incomplete gamma function. The prior distribution is described by $f(\lambda)$ with parameters a' and b'. The parameters of the posterior $f(\lambda|x,t)$ are denoted a'' and b'' and are obtained as:

$$a'' = a' + x \tag{21}$$

$$b'' = b' + t \tag{22}$$

Inserting λ_{SL} together with a'' and b'' into Eq. (20) provides the answer to the key question in this probabilistic reliability assessment: What is the probability $\Pr(\lambda < \lambda_{SL}|x,t)$ that the system under consideration complies with the target level of safety λ_{SL}?

$$\Pr(\lambda < \lambda_{SL}|x,t) = F(\lambda_{SL}|x,t) \tag{23}$$

Moreover, the best point estimate of the unknown error rate λ is the posterior mean $\hat{\lambda}$:

$$\hat{\lambda} = \frac{a''}{b''} = \frac{a' + x}{b' + t} \tag{24}$$

As the analysis deals with a safety-relevant issue, often a more conservative estimate for λ than the posterior mean $\hat{\lambda}$ is desired. Therefore the analyst may chose for instance the 95 % quantile, or alternatively the 99 % quantile, of the posterior λ as a conservative point estimate.

To perform the analysis, prior parameters have to be selected. A commonly accepted formal rule to construct an (objective) prior distribution when no prior information is available has been defined by Jeffreys [40–42]. The property that makes Jeffreys' prior non-informative is its invariance to re-parameterizations [40]. Here, Jeffreys' prior yields $a' = 0.5$ and $b' \to 0$. Eq. (24) supports the interpretation of the prior parameters as a' prior error observations in a prior test time interval b' (see [21] page 89). However, by comparing Eq. (19) with Eq. (8), Gelman et al. [40] page 52 argue that the prior parameters may be interpreted as $a' - 1$ prior observations in a prior time interval b'. Following this interpretation, a weakly informative prior in case no prior information is available could also be selected as $a' = 1$ and $b' \to 0$. One is able to show that with $a' = 1$ and $b' \to 0$ the same numerical results for the necessary test effort t are obtained as with NHST. If substantial information prior to the analysis is available, then this information can easily be incorporated into a' and b' following the interpretation given.

The test drive effort before collecting the data can be derived by the following steps:

- Select the probability with which the target level of safety should be complied with (e.g. $\Pr(\lambda < \lambda_{SL}|x,t) = 0.95$)
- Insert the desired target level of safety λ_{SL} into Eq. (20)
- Fix the acceptable number of errors x of the test drive at different values (i.e. $x = 0,1,2,\ldots$)
- For each value of x, solve Eq. (20) for the test effort t

The result are the acceptable number of errors x for a given test drive effort t, which all allow to conclude with at least 95 % probability that the sensor complies with λ_{SL}.

Assessing the reliability of a multi-sensor system

To enhance the safety of the environment perception, the system architecture may include redundant sensors obtaining the same types of information in overlapping field of views. It is pointed out that complementary sensor principles used to obtain different types of information (e.g. camera for object classification and radar for object localization) do not comprise a redundant but a complementary system [37]. These are not considered here.

To combine the information and data of multiple sensors, sensor data fusion is applied [43]. Modern sensor data fusion algorithms mostly are based on Bayes filters such as the well-known Kalman filter [43]. Because it is not straightforward to evaluate the performance of a multi-sensor system with complicated fusion algorithms, we simplify the problem with a so-called majority voting system [12, 15]. The assumption is that a system's perception using the information of redundant sensors fails, when more than half of the individual sensors commit safety-relevant perception errors. An approach utilizing a majority voting scheme to describe the multi-sensor based perception reliability has already been reported in [15].

In reliability analysis, a majority voting system can be represented as a k-out-of-N system [22], meaning that at least k-out-of-N sensors have to commit safety-relevant perception errors for the system to provide erroneous information. To calculate the multi-sensor system's rate of perception error occurrence, let the occurrence of safety-relevant errors for each sensor $s = 1, \dots, N$ (N is the total number of redundant sensors) for a given measurement cycle be a binary random variable U_s with $u_s = 1$ meaning sensor s commits at least one error and $u_s = 0$ meaning sensor s commits no error. The probability p of committing an error is assumed to be equal for all sensors and is related to the error rate λ of the individual sensors as well as the measurement cycle time t_{cycle}:

$$p = 1 - \exp(-\lambda \cdot t_{cycle}) \approx \lambda \cdot t_{cycle} \tag{25}$$

The approximation $p = \lambda \cdot t_{cycle}$ holds for $\lambda \ll 1\ \mathrm{h}^{-1}$. The multi-sensor machine vision, based on majority voting, commits perception errors when $\sum_{s=1}^{N} U_s \geq \left\lfloor \frac{N}{2} + 1 \right\rfloor$, with $\left\lfloor \frac{N}{2} + 1 \right\rfloor$ being the notation for rounding $\frac{N}{2} + 1$ down. Under the assumption that the individual sensors' perception error probabilities p are independent of each other, the probability p_f of the multi-sensor based machine vision to provide erroneous information can be calculated with the binomial CDF:

$$p_f = \sum_{k=\left\lfloor \frac{N}{2}+1 \right\rfloor}^{N} \binom{N}{k} \cdot p^k \cdot (1-p)^{N-k} \tag{26}$$

Adverse physical conditions such as the presence of objects with low reflectivity or strong rain gusts might lead to dependence between potential safety-relevant perception errors of redundant sensors (equivalent to the discussion in the previous section). Therefore the assumption of independence in Eq. (26) might not be justified. To take dependent multi-sensor errors into account, we define the correlation coefficient ρ of perception error occurrence U_s and U_q between any pairs of sensors $s, q \,\epsilon\, \{1, \dots, N\}$ [44]:

$$\rho = \frac{\mathrm{E}[U_s \cdot U_q] - \mathrm{E}[U_s] \cdot \mathrm{E}[U_q]}{\sqrt{\mathrm{E}[U_s](1 - \mathrm{E}[U_s]) \cdot \mathrm{E}[U_q] \cdot (1 - \mathrm{E}[U_q])}} \tag{27}$$

where E[ ] denotes the expectation operator. An important aspect is the interpretation of the correlation coefficient given by Eq. (27), which in this form is not very intuitive. When identical sensors are utilized, it holds $\mathrm{E}[U_s] = \mathrm{E}[U_q] = p$ and further:

$$\begin{aligned} \mathrm{E}[U_s \cdot U_q] &= \Pr(U_s = 1 \cap U_q = 1) = \cdots \\ &= \Pr(U_s = 1 | U_q = 1) \cdot \Pr(U_q = 1) = \cdots \\ &= \Pr(U_s = 1 | U_q = 1) \cdot p \end{aligned} \tag{28}$$

Inserting into Eq. (27) leads to:

$$\rho = \frac{\Pr(U_s = 1 | U_q = 1) \cdot p - p^2}{p - p^2} \tag{29}$$

Eq. (29) is now more easily interpreted: $\Pr(U_s = 1 | U_q = 1)$ is the conditional probability that sensor s commits a perception error given that sensor q has committed an error. p is the individual sensors' probability of perception error occurrence. On the one hand, with independence, i.e. $\Pr(U_s = 1 | U_q = 1) = \Pr(U_s = 1) = p$, the correlation coefficient becomes $\rho = 0$. On the other hand, if it is certain that sensor s commits an error when an error occurs in sensor q, i.e. $\Pr(U_s = 1 | U_q = 1) = 1$, the correlation coefficient becomes $\rho = 1$. This is equivalent to full dependence. Finally, when p is small, it holds $p^2 \ll p$ and Eq. (29) can be simplified to:

$$\rho \approx \Pr(U_s = 1 | U_q = 1) \tag{30}$$

That is, the correlation coefficient ρ is approximately equal to the conditional probability of error occurrence in sensor s given an error has occurred in sensor q.

To account for perception error dependence between redundant sensors we consider the beta-binomial distribution [44–47] and a model for correlated binary data proposed by Gupta and Tao [48]. The latter model is introduced because the beta-binomial distribution can due to numerical reasons not be utilized with small values of the correlation coefficient ρ. Conversely, the Gupta and Tao model is not applicable to large values of correlation ρ (for further information see [49, 50]). The exact values of ρ up to which both models can be used depend on the probability p. In the subsequent case study we utilize the beta binomial model for $\rho \geq 0.01$ and the Gupta and Tao model for $\rho < 0.01$.

With the beta-binomial distribution, the probability $\Pr(\sum_{s=1}^{N} U_s = k)$ of exactly k-out-of-N sensors committing perception errors is [46]:

$$\Pr\left(\sum_{s=1}^{N} U_s = k\right) = \binom{N}{k} \frac{\Gamma(\theta_1 + \theta_2) \cdot \Gamma(\theta_1 + k) \cdot \Gamma(\theta_2 + N - k)}{\Gamma(\theta_1) \cdot \Gamma(\theta_2) \cdot \Gamma(\theta_1 + \theta_2 + \mathrm{N})} \tag{31}$$

with $\Gamma(\,)$ the gamma function and the parameters θ_1, θ_2 related to the individual sensors' probability of perception error occurrence p and the correlation coefficient ρ (derived from [47]):

$$\theta_1 = \frac{p \cdot (1 - \rho)}{\rho}, \quad \theta_2 = \frac{(1 - p) \cdot (1 - \rho)}{\rho} \tag{32}$$

Both p and ρ are assumed to be the equal for all sensors $s = 1, \dots, N$ and p is given by Eq. (25) when the error rate λ of an individual sensor is known. The probability of the majority vote based multi-sensor machine vision to commit a perception error p_f, accounting for dependent sensor errors, can then be calculated as:

$$p_f = \sum_{k=\left\lfloor \frac{N}{2}+1 \right\rfloor}^{N} \Pr\left(\sum_{s=1}^{N} U_s = k\right) \tag{33}$$

Eq. (25), (31) and (33) allow to study the system's perception reliability including error dependence among redundant sensors according to the beta-binomial model. The solution utilizing the Gupta and Tao [48] model is given in the appendix.

In theory, when the correlation coefficient goes to $\rho \rightarrow 0$, both the beta-binomial distribution as well as the Gupta and Tao model converge to the (independent) binomial distribution [46, 48]. Thus with $\rho \rightarrow 0$ the multi-sensor probability of perception error occurrence converges to Eq. (26), independently of which of the two models is utilized.

Case study: Empirically demonstrating the perception reliability of environment sensors

A synthetic case study is performed. Suppose one is interested in demonstrating that a sensor's environment perception fulfills the target level of safety $\lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$. For this target, first the necessary test drive effort t to demonstrate $\lambda < \lambda_{SL}$ is derived. With the derived test effort it is demonstrated how the test drive has to be conducted to account for different weather influences. It is assumed that the weather in a particular region can be described with the probabilities $p_{sun} = 0.65$; $p_{rain} = 0.15$; $p_{snow} = 0.05$ and $p_{cloudy} = 0.15$.

Thereafter, we draw probabilistic conclusions about the safety-relevant perception error rate λ based on hypothetical results of a test drive with the aim of illustrating the uncertainty in estimating λ.

In the last section of the case study, we study the influence of dependence on perception error occurrence in a multi-sensor based machine vision system. To this end, we assume that a specific task of the machine vision (e.g. object detection in the front field of view) is based on three identical Lidar or Radar sensors, respectively. The individual error rate λ of the three sensors is thus identical. The measurement cycle time is assumed to be $t_{cycle} = 0.05$ s.

For all calculations in this case study, Jeffreys' prior parameters $a' = 0.5$ and $b' = 0$ are selected in Eq. (18). Jeffreys' prior reflects ignorance on the error rate λ before conducting the test and is commonly considered to be non-informative [40]. This choice of prior is conservative as it only assigns a prior probability of $\Pr(\lambda < \lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}) = 1.13 \cdot 10^{-11}$ that the target level of safety is met.

Estimating the necessary test drive effort

The empirical test drive effort can be estimated with Eq. (23). Figure 3a) shows the probability $\Pr(\lambda < \lambda_{SL}|x,t)$ that the sensor under consideration complies with the target level of safety $\lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$ for the cases of $x = 0$, $x = 1$ and $x = 2$ perception errors during a test drive, in function of the test drive effort t. To derive a conservative estimate of the test drive effort t required for a demonstration of $\lambda < \lambda_{SL}$, one may select the t for which it holds $\Pr(\lambda < \lambda_{SL}|x,t) = 0.95$ as indicated for the case $x = 0$ in Figure 3a) with the grey arrow.

A summary of different combinations of test drive efforts t and the corresponding acceptable number of errors x that all allow to conclude $\Pr(\lambda < \lambda_{SL}|x,t) = 0.95$ are given in Figure 3b). The smallest possible test effort to demonstrate $\lambda < \lambda_{SL}$ requires the observation of no errors and is $t = 1.92 \cdot 10^{7}$ h. Even though all combinations in Figure 3b) show compliance with the target level of safety with 95% credibility, it is necessary to select the test drive effort a-priori, and not do adjust it based on the observed number of errors. If a stopping criteria is selected on the go, the estimate of $\Pr(\lambda < \lambda_{SL}|x,t)$ is biased.

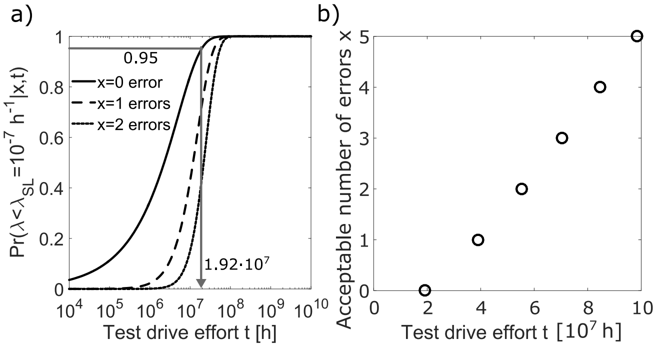

Figure 3. a) Probability $\Pr(\lambda < \lambda_{SL}|x,t)$ of compliance with the target level of safety $\lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$ in function of the test drive effort t, for the cases of $x = 0$, $x = 1$ and $x = 2$ errors during the test. The grey arrow indicates the test drive effort for the case of $x = 0$ errors and $\Pr(\lambda < \lambda_{SL}|x,t) = 0.95$.
b) Number of acceptable errors x in a test drive as a function of the test drive effort t such that $\Pr(\lambda < \lambda_{SL}|x,t) = 0.95$.

Suppose the test effort is selected as $t = 1.92 \cdot 10^{7}$ h with no acceptable errors x during the test. To consider that error rates might differ substantially in function of environmental conditions with influence on sensor performance, the test effort has to be distributed according to the probabilities of the different environment conditions. Here only weather conditions are considered as relevant influencing factors. The resulting test profile is summarized in Table 1. Application of this test profile takes non-stationary error rates according to Eq. (17) into account.

Table 1. Resulting test profile to account for different weather conditions.

Weather condition	**Test time t_i for weather condition i**
Sunny	$t_{sun} = 1.248 \cdot 10^{7}$ h
Rain	$t_{rain} = 0.288 \cdot 10^{7}$ h
Snow	$t_{snow} = 0.096 \cdot 10^{7}$ h
Cloudy	$t_{cloudy} = 0.288 \cdot 10^{7}$ h

Evaluating hypothetical test results

In this section it is assumed that a test drive with $t = 1.92 \cdot 10^{7}$ h has been conducted. Two hypothetical results of this test drive are evaluated: (a) $x = 0$ and (b) $x = 1$ errors have been observed in t. With these test results, the posterior parameters of the gamma distribution are calculated according to Eq. (21) and Eq. (22): (a) $a'' = 0.5$, $b'' = 1.92 \cdot 10^{7}$ and (b) $a'' = 1.5$, $b'' = 1.92 \cdot 10^{7}$. The resulting posterior PDFs and CDFs of the error rate λ with these parameters are illustrated in Figure 4.

As the PDFs in Figure 4a) and Figure 4c) show, the Bayesian approach captures the full uncertainty in the error rate λ by assigning to each value of the error rate a probability density. The CDF Figure 4b) shows for $x = 0$ that the unknown error rate is with 95% probability smaller than $\lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$. This is in accordance with the test design derived in the previous section. With the observation of $x = 1$ error, the error rate is with 95% probability

$\lambda < 2 \cdot 10^{-7}\ \mathrm{h}^{-1}$, as illustrated in Figure 4d). The target level of safety $\lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$ is thus not fulfilled on basis of the 95% quantile.

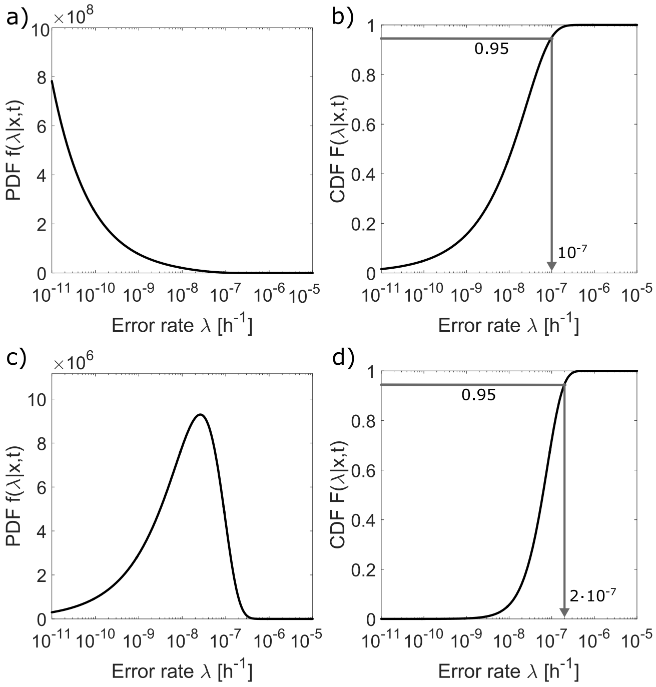

Figure 4. a) Posterior PDF $f(\lambda|x,t)$ of the error rate when observing x=0 errors in time $t = 1.92 \cdot 10^7$ h and b) corresponding CDF. c) Posterior PDF $f(\lambda|x,t)$ of the rate when observing x=1 error in time $t = 1.92 \cdot 10^7$ h and d) corresponding CDF. The grey arrows indicate the 95 % quantiles of the error rate λ.

Influence of error dependence on multi-sensor based machine vision

While the previous two sections dealt with the reliability assessment of an individual sensor, this section studies the reliability of a multi-sensor system where we model the sensor data fusion with a majority voting scheme. The multi-sensor system consists of three identical redundant sensors. First it is assumed that all sensors comply with the target level of safety such that each individual sensor has a perception error rate of exactly $\lambda = \lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$. However, the correlation coefficient of error occurrence ρ between the three sensors is unknown. What is now the error rate λ_{system} of the multi-sensor based machine vision? The answer to this question can be obtained with Eqs. (31) to (33) for the beta-binomial distribution and with Eqs. (A1), (A2) and Eq. (33) for the Gupta and Tao model [48].

λ_{system} is illustrated in Figure 5 in function of the correlation coefficient ρ in semi- and in double-logarithmic scale. The semi-logarithmic plot in Figure 5 shows that the Gupta and Tao model (dashed line) in this specific case cannot be utilized for $\rho > 0.5$ and deviates from the beta-binomial distribution starting around $\rho = 0.05$. In the range of $\rho > 0.01$, we therefore utilize the beta-binomial model that converges against the individual sensors' perception error rate $\lambda = 10^{-7}\ \mathrm{h}^{-1}$ with full dependence ($\rho \to 1$). This is the intuitive solution of a fully dependent redundant system. The beta-binomial model (solid line) cannot be evaluated for $\rho < 0.006$ due to numerical reasons. Thus, in the range $\rho < 0.01$ the Gupta-Tao model is utilized, which converges against the solution of the independent binomial CDF Eq. (26) with $\rho \to 0$. With independence, the system's error rate is $\lambda_{system} = 4.2 \cdot 10^{-19}\ \mathrm{h}^{-1}$.

In Figure 5 it is visible that for $\rho < 0.1$ the system's error rate $\lambda_{\mathrm{system}}$ strongly depends on the correlation coefficient. The reason for this sensitivity is explained with the interpretation of the correlation coefficient given through Eq. (29) and (30). For instance if $\rho \approx Pr\left(U_s = 1 | U_q = 1\right) = 10^{-5}$, then the conditional probability of perception error occurrence in sensor s – given an error has occurred in sensor q – is $7.2 \cdot 10^6$ times larger than in the independent case. As can be seen in Figure 5b, this leads to a system error rate $\lambda_{\mathrm{system}} \approx 3 \cdot 10^{-12}\ \mathrm{h}^{-1}$, substantially larger than with independent component errors.

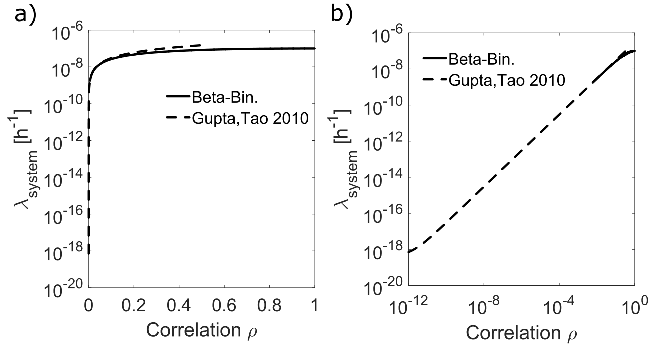

Figure 5. a) Perception error rate λ_{system} of a redundant multi-sensor based machine vision system in dependence of the correlation coefficient ρ. The system consists of three identical sensors with $\lambda = \lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$ each. The solid line represents the beta-binomial model and the dashed line the model presented in [48] (see appendix). b) The same plot in double logarithmic scale.

In the previous sections of this case study, the test drive effort was derived such that an individual sensor complies with $\lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$. However, the relationship between the system's error rate λ_{system} and correlation among the sensors has implications on the test effort, when the target level of safety is set on the system level. The two important questions then are: How large does the perception error rate of an individual sensor has to be and how much test drive effort is necessary at the individual sensor level, such that the system complies with the target level of safety $\lambda_{system} \leq \lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$? The solution to these questions is illustrated in Figure 6 in function of the error correlation among different sensors.

For clarity, the target level of safety for the individual sensors in a redundant multi-sensor system that lead to an overall system perception error rate of $\lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$ is denoted with $\lambda_{individual,SL}$ and is shown in Figure 6a. Figure 6b shows the corresponding test drive effort which is necessary to demonstrate that the target level of safety of an individual sensor $\lambda_{individual,SL}$ is complied with $\mathrm{Pr}\left(\lambda \leq \lambda_{individual,SL} | x, t\right) = 0.95$. If all sensors are independent of each other, the individual safety target is as low as $\lambda_{individual,SL} = 0.05\ \mathrm{h}^{-1}$, only requiring a test drive effort of $t = 40$ h to demonstrate that the system complies with the target level of safety, i.e. $\mathrm{Pr}\left(\lambda \leq \lambda_{individual,SL} | x, t\right) = \mathrm{Pr}\left(\lambda_{system} \leq \lambda_{SL}\right) = 0.95$. In contrast, if all sensors are fully dependent, the safety target of the individual sensors reduces to $\lambda_{individual,SL} = \lambda_{SL} = 10^{-7}\ \mathrm{h}^{-1}$ with a test drive effort of $t = 1.92 \cdot 10^7$ h. Obviously this is the same test drive effort as when implementing only a single sensor. With a correlation coefficient of $\rho \leq 10^{-4}$ the test drive effort is $t \leq 5.9 \cdot 10^3$ h.

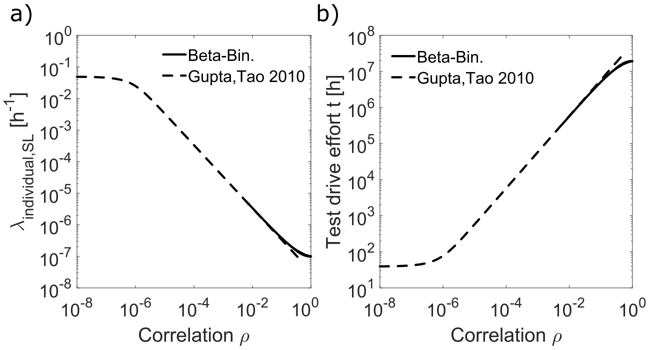

Figure 6. a) Necessary target level of safety for the individual sensors $\lambda_{individual,SL}$ in a redundant multi-sensor system (3 identical sensors) such that the system complies with $\lambda_{system} \leq \lambda_{SL} = 10^{-7}$ h^{-1}. b) Corresponding test drive effort t for $\Pr\left(\lambda \leq \lambda_{individual,SL} | x, t\right) = 0.95$ when no error is accepted in t. Both a) and b) are in function of the correlation coefficient ρ of error occurrence among the different sensors.

Discussion

The interest is in the test drive effort that allows an empirical compliance demonstration of a sensors' environment perception with the (for demonstrative purposes selected) target level of safety $\lambda_{SL} = 10^{-7}$ h^{-1}. For individual sensors, this test drive effort is found to be in the order of at least $t = 1.92 \cdot 10^{7}$ h, which appears infeasible in most practical contexts. Similar conclusions have already been drawn in [14], in which the test drive effort to demonstrate an automated system's safety is derived with the Null Hypothesis Significance Testing (NHST). This indicates that the empirical demonstration (i.e. test drives in real driving situations) might not be the way to show that a sensor's perception is sufficiently safe when the target level of safety is strict.

In case of less restrictive safety requirements however, the here presented Bayesian approach allows to estimate test efforts. It has two main advantages over NHST: 1) it provides results that are easy to interpret, and 2) it is more flexible in extending the model, in particular to a hierarchical model that addresses changing environmental factors, and to a multi-sensor system. Figure 4 shows how the presented Bayesian approach captures the full uncertainty in estimating the perception error rate λ, which is more intuitively understood than statements on the statistical significance of a hypothesis about λ. The graphical illustrations in Figure 4 and Figure 1 underline that the Bayesian methodology can readily be interpreted as a probability, while the p-value and significance level α of NHST are more difficult to understand (and are not understood by most engineers). Due to its intuitive and easy interpretation, the communication with decision makers benefits from the Bayesian approach: Given that the statistical model and assumptions represent the problem adequately well, the Bayesian test methodology results in the probability $\Pr(\lambda < \lambda_{SL} | x, t)$ that for a specific set of observations the target level of safety is complied with.

In case certain aspects of the environment perception such as object localization are based on multiple redundant sensors, the redundancy should be considered in the reliability assessment of environment perception. As in every redundant system, redundancy can drastically increase the system reliability if sensors perform independently. However, Figure 5 should serve as a warning not to assume independence light handed, as the true error rate of the perception system might then be underestimated. For the investigated system, a seemingly small correlation of $\rho = 10^{-5}$ increases the system's perception error rate by a factor of $7.1 \cdot 10^{6}$ compared to the case of independence ($\rho = 0$)!

It is illustrated in Figure 6 that the overall target level of safety λ_{SL} of the machine vision may be more easily demonstrated in a redundant multi-sensor system than for an individual sensor, as long as errors dependence among the different sensors is small. In the extreme case of error independence at different sensors ($\rho = 0$), a system consisting of three sensors would require a test effort of only $t = 40$ h to comply with $\lambda_{SL} = 10^{-7}$ h^{-1}. This illustrates how a redundant sensor system may offer the opportunity to demonstrate the safety of the environment perception with economical feasible effort in an empirical way, i.e. test drives with individual sensors in real driving situations. This result may be especially relevant for the future when the costs of environment sensors decrease. For instance, the reliability of the environment perception could be more easily and with lower costs ensured by multiple (independent) mid-class sensors than with one high-end sensor. However, in order to evaluate whether this is a valid alternative, one would need to know the correlation ρ of error occurrence among different sensors. The problem therefore shifts to demonstrating a low correlation ρ, which likely will require a larger test effort. This study should therefore be extended to determine the test effort necessary to determine a sufficiently low correlation between the sensors.

It is important to note that the presented approach does not model the sensor data fusion in its full complexity. Instead of analyzing real fusion algorithms, the results are here derived by simplifying the sensor data fusion with a majority decision. While majority voting is a valid method to increase a system's reliability [12], it is doubtful whether this method finds widespread application in practice. It is likely that modern fusion algorithms [43] will outperform a majority voting system, therefore the presented results may be regarded as conservative.

The presented study combines different types of perception errors into the single metric "error rate λ". In reality, different types of perception errors include false-positive and false-negative object detections, errors in physical measurement quantities and object classification errors. The different types of errors occur with varying probability and may not be equally safety-relevant. It is here entirely left to the analyst to decide what comprises a safety-relevant perception error.

Another limitation is found in the treatment of the temporal variability of error occurrence, induced through different physical influencing conditions. In this contribution we do not answer the question of how to assess which factors with influence on sensor performance to consider, this has to be decided by experts and preliminary test results. Also, as stated before, the probability of the different influencing factors such as weather is dependent on the geographical region. Future work should optimize the presented work to account for this aspect and examine how to include different environment conditions in higher detail.

Conclusions

A Bayesian methodology for empirical reliability assessments of sensor based environment perception is presented as an alternative to the commonly applied Null Hypothesis Significance Testing (NHST). It allows to estimate the necessary test drive effort to demonstrate the perception reliability of environment sensors, including dependent errors and time variable error probabilities. Furthermore, a solution to assess the reliability of a dependent redundant multi-sensor system is given.

Applying the methodology in a case study shows that the empirical test drive effort may be unfeasibly large when the target level of safety is low. When working with a multi-sensor system in which the individual sensors are nearly independent of each other, the system's perception reliability is considerably higher than when utilizing a single sensor. This fact opens up the possibility of validating the perception reliability empirically with feasible test drive effort, when one is able to show that multiple sensors have a small error dependency. The verification of a small error dependency itself is however expected to require additional test drive efforts. It remains to be studied what the necessary test setup and effort is for this purpose. Simplifications of the problem's complexity involve the treatment of different types of perception errors, the representation of the sensor data fusion with a majority voting scheme and in approximating the time dependent performance of the perception induced through various physical influencing factors such as the weather.

References

1. Beiker, S., "Deployment Scenarios for Vehicles with Higher-Order Automation," in: Maurer, M., Gerdes, J.C., Lenz, B., and Winner, H. (eds.), *Autonomous Driving*, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-662-48845-4:193–211, 2016.
2. Google, "Google Self-Driving Car Project," http://static.googleusercontent.com/media/www.google.com/en/us/selfdrivingcar/, February 25, 2016.
3. Audi, "Mission accomplished: Audi A7 piloted driving car completes 550-mile automated test drive," https://www.audiusa.com/newsroom/news/press-releases/2015/01/550-mile-piloted-drive-from-silicon-valley-to-las-vegas, February 25, 2016.
4. Pink, O., Becker, J., and Kammel, S., "Automated driving on public roads: Experiences in real traffic," *it - Information Technology* 57(4), 2015, doi:10.1515/itit-2015-0010.
5. Broggi, A., Debattisti, S., Grisleri, P., and Panciroli, M., "The deeva autonomous vehicle platform," *2015 IEEE Intelligent Vehicles Symposium (IV)*, Seoul, South Korea:692–699.
6. Aeberhard, M., Rauch, S., Bahram, M., Tanzmeister, G. et al., "Experience, Results and Lessons Learned from Automated Driving on Germany's Highways," *IEEE Intell. Transport. Syst. Mag.* 7(1):42–57, 2015, doi:10.1109/MITS.2014.2360306.
7. Franke, U., Pfeiffer, D., Rabe, C., Knoeppel, C. et al., "Making Bertha See," *2013 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Sydney, Australia:214–221, 2013.
8. Kammel, S., Ziegler, J., Pitzer, B., Werling, M. et al., "Team AnnieWAY's autonomous system for the 2007 DARPA Urban Challenge," *J. Field Robotics* 25(9):615–639, 2008, doi:10.1002/rob.20252.
9. Blevis, B., "Losses due to rain on radomes and antenna reflecting surfaces," *IEEE Trans. Antennas Propagat.* 13(1):175–176, 1965, doi:10.1109/TAP.1965.1138384.
10. Ishimaru, A., "Wave propagation and scattering in random media and rough surfaces," *Proc. IEEE* 79(10):1359–1366, 1991, doi:10.1109/5.104210.
11. Rasshofer, R.H., Spies, M., and Spies, H., "Influences of weather phenomena on automotive laser radar systems," *Adv. Radio Sci.* 9:49–60, 2011, doi:10.5194/ars-9-49-2011.
12. Weitzel, A., Winner, H., Peng, C., Geyer, S. et al., "Absicherungsstrategien für Fahrerassistenzsysteme mit Umfeldwahrnehmung: [Bericht zum Forschungsprojekt FE 82.0546/2012]," Berichte der Bundesanstalt für Straßenwesen Fahrzeugtechnik, vol. 98, Fachverl. NW, Bremen, ISBN 978-3-95606-118-9, 2014.
13. Wachenfeld, W. and Winner, H., "The Release of Autonomous Vehicles," in: Maurer, M., Gerdes, J.C., Lenz, B., and Winner, H. (eds.), *Autonomous Driving*, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-662-48845-4:425–449, 2016.
14. Winner, H., "ADAS, Quo Vadis?," in: Winner, H., Hakuli, S., Lotz, F., and Singer, C. (eds.), *Handbook of Driver Assistance Systems*, Springer International Publishing, Cham, ISBN 978-3-319-12351-6:1557–1584, 2016.
15. Bock, F., Siegl, S., and German, R., "Mathematical Test Effort Estimation for Dependability Assessment of Sensor-based Driver Assistance Systems," *2016 42st Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Limassol, Cyprus.
16. Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B. et al., "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations," *European journal of epidemiology* 31(4):337–350, 2016, doi:10.1007/s10654-016-0149-3.
17. Cohen, J., "The earth is round (p < .05)," *American Psychologist* 49(12):997–1003, 1994, doi:10.1037/0003-066X.49.12.997.
18. Nuzzo, R., "Scientific method: statistical errors," *Nature* 506:150–152, 2014, doi:10.1038/506150a.
19. Baker, M., "Statisticians issue warning over misuse of P values," *Nature* 531:151, 2016, doi:10.1038/nature.2016.19503.
20. Wasserstein, R.L. and Lazar, N.A., "The ASA's Statement on p - Values: Context, Process, and Purpose," *The American Statistician* 70(2):129–133, 2016, doi:10.1080/00031305.2016.1154108.
21. Hamada, M.S., Wilson, A.G., Reese, C.S., and Martz, H.F., "Bayesian Reliability," Springer Series in Statistics, 1st ed., Springer-Verlag, s.l., ISBN 978-0-387-77950-8, 2008.
22. Rausand, M. and Høyland, A., "System reliability theory: Models, statistical methods, and applications," Wiley series in probability and statistics Applied probability and statistics, 2nd ed., Wiley-Interscience, Hoboken, NJ, ISBN 978-0-471-47133-2, 2004.
23. Fitzgerald, M., Martz, H.F., and Parker, R.L., "Bayesian Single-Level Binomial And Exponential Reliability Demonstration Test Plans," *Int. J. Rel. Qual. Saf. Eng.* 06(02):123–137, 1999, doi:10.1142/S0218539399000139.
24. Singh, H., Cortellessa, V., Cukic, B., Gunel, E. et al., "A Bayesian approach to reliability prediction and assessment of component based systems," *12th International Symposium on Software Reliability Engineering. ISSRE 2001*, Hong Kong, China, 27-30 Nov. 2001:12–21.
25. Brender, D.M., "The Bayesian Assessment of System Availability: Advanced Applications and Techniques," *IEEE Trans. Rel.* R-17(3):138–147, 1968, doi:10.1109/TR.1968.5216927.
26. Martz, H.F. and Wailer, R.A., "Bayesian reliability analysis of complex series/parallel systems of binomial subsystems and components," *Technometrics* 32(4):407–416, 1990.
27. Guida, M. and Pulcini, G., "Automotive reliability inference based on past data and technical knowledge," *Reliability Engineering & System Safety* 76(2):129–137, 2002, doi:10.1016/S0951-8320(01)00132-6.
28. Gotzig, H. and Geduld, G., "Automotive LIDAR," in: Winner, H., Hakuli, S., Lotz, F., and Singer, C. (eds.), *Handbook of Driver Assistance Systems*, Springer International Publishing, Cham, ISBN 978-3-319-12351-6:405–430, 2016.
29. Winner, H., "Automotive RADAR," in: Winner, H., Hakuli, S., Lotz, F., and Singer, C. (eds.), *Handbook of Driver Assistance Systems*, Springer International Publishing, Cham, ISBN 978-3-319-12351-6:325–403, 2016.

30. Dietmayer, K., “Predicting of Machine Perception for Automated Driving,” in: Maurer, M., Gerdes, J.C., Lenz, B., and Winner, H. (eds.), *Autonomous Driving*, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-662-48845-4:407–424, 2016.
31. Kottegoda, N.T. and Rosso, R., “Applied statistics for civil and environmental engineers,” 2nd ed., Blackwell Pub, Oxford, UK, [Malden, MA], ISBN 978-1-4051-7917-1, 2008.
32. Vidgen, B. and Yasseri, T., “P-Values: Misunderstood and Misused,” *Front. Phys.* 4:e124, 2016, doi:10.3389/fphy.2016.00006.
33. Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D. et al., “The fallacy of placing confidence in confidence intervals,” *Psychonomic bulletin & review* 23(1):103–123, 2016, doi:10.3758/s13423-015-0947-8.
34. VanderPlas, J., “Frequentism and bayesianism: a python-driven primer,” *arXiv preprint arXiv:1411.5018*, 2014.
35. Jaynes, E.T. and Kempthorne, O., “Confidence Intervals vs Bayesian Intervals,” in: Harper, W.L. and Hooker, C.A. (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Springer Netherlands, Dordrecht, ISBN 978-90-277-0619-5:175–257, 1976.
36. Gelman, A., “Induction and deduction in Bayesian data analysis,” *Rationality, Markets and Morals* 2:67–78, 2011.
37. Darms, M., “Data Fusion of Environment-Perception Sensors for ADAS,” in: Winner, H., Hakuli, S., Lotz, F., and Singer, C. (eds.), *Handbook of Driver Assistance Systems*, Springer International Publishing, Cham, ISBN 978-3-319-12351-6:549–566, 2016.
38. Winner, H., “Fundamentals of Collision Protection Systems,” in: Winner, H., Hakuli, S., Lotz, F., and Singer, C. (eds.), *Handbook of Driver Assistance Systems*, Springer International Publishing, Cham, ISBN 978-3-319-12351-6:1149–1176, 2016.
39. Straub, D., “Lecture Notes in Engineering Risk Analysis,” 2013.
40. Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., “Bayesian data analysis,” Texts in statistical science, 2nd ed., Chapman & Hall, Boca Raton, Fla., ISBN 9781584883883, 2004.
41. Jeffreys, H., “Theory of Probability,” 3rd ed., Oxford University press, London, 1961.
42. Kass, R.E. and Wasserman, L., “The selection of prior distributions by formal rules,” *Journal of the American Statistical Association* 91(435):1343–1370, 1996.
43. Durrant-Whyte, H. and Henderson, T.C., “Multisensor Data Fusion,” in: Siciliano, B. and Khatib, O. (eds.), *Springer Handbook of Robotics*, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-23957-4:585–610, 2008.
44. Hisakado, M., Kitsukawa, K., and Mori, S., “Correlated binomial models and correlation structures,” *Journal of Physics A: Mathematical and General* 39(50):15365, 2006.
45. Rosner, B., “Beta-Binomial Distribution,” in: Armitage, P. and Colton, T. (eds.), *Encyclopedia of Biostatistics*, John Wiley & Sons, Ltd, Chichester, UK, ISBN 047084907X, 2005.
46. Maharry, T.J., “PROPORTION DIFFERENCES USINGTHE BETA-BINOMIAL DISTRIBUTION,” Dissertation, Oklahoma State University, Stillwater, Oklahoma, 2006.
47. Paul, S.R., “Applications of the beta distribution,” in: Gupta, A.K. and Nadarajah, S. (eds.), *Handbook of beta distribution and its applications*, CRC press:423–436, 2004.
48. Gupta, R.C. and Tao, H., “A generalized correlated binomial distribution with application in multiple testing problems,” *Metrika* 71(1):59–77, 2010, doi:10.1007/s00184-008-0202-7.
49. Bahadur, R.R., “A representation of the joint distribution of responses to n dichotomous items,” *Studies in item analysis and prediction* 6:158–168, 1961.
50. Kupper, L.L. and Haseman, J.K., “The use of a correlated binomial model for the analysis of certain toxicological experiments,” *Biometrics*:69–76, 1978.

Contact Information

Mario Berk, Ph.D. student at the Engineering Risk Analysis Group, Technical University of Munich in cooperation with AUDI AG, INI.TUM

E-Mail: mario.berk@tum.de

Daniel Straub, Professor of Engineering Risk Analysis, Technical University of Munich

E-Mail: straub@tum.de

Appendix

To complement the results of the beta-binomial distribution, the correlated binomial distribution suggested in Gupta, Tao [48] is used. With this model the probability $\Pr(\sum_{s=1}^{N} U_s = k)$ of exactly k-out-of-N sensors to commit a safety-relevant error, assuming each sensor has the same perception error probability p, is:

$$\Pr\left(\sum_{s=1}^{N} U_s = k\right) = p \cdot \Pr\left(\sum_{s=1}^{N-1} U_s = k-1\right) + (1-p) \cdot \Pr\left(\sum_{s=1}^{N-1} U_s = k\right) + \rho \cdot \sum_{s=1}^{N-1} p(1-p) \cdot a_{N,k}^{s} \tag{A1}$$

Where ρ is the correlation coefficient quantifying the correlation of error occurrence between the sensors. It is assumed ρ is equal among all sensors $s = 1, \dots, N$. The factor $a_{N,k}^{s}$ is defined as:

$$a_{N,k}^{s} = \begin{cases} 0, & if\ k < 0\ or\ k > N \\ a_{2,0}^{1} = 1, a_{2,1}^{1} = -2, a_{2,2}^{1} = 1, & if\ N = 2,\ s = 1 \\ p \cdot a_{N-1,k-1}^{s-1} + (1-p) a_{N-1,k}^{s-1}, & if\ N > 2,\ s = N-1 \\ p \cdot a_{N-1,k-1}^{s} + (1-p) a_{N-1,k}^{s}, & if\ N > 2,\ s = 1,2,\dots,N-2 \end{cases} \tag{A2}$$

Inserting Eq. (A1) into Eq. (33) yields the probability of the majority vote based multi-sensor machine vision to fail p_f, including dependence according to the Gupta, Tao model.