Towards an Advanced Self-Monitoring Tracking Module: Leveraging Statistical Hypothesis Tests and Subjective Logic Reasoning

Thomas Griebel, Alexander Scheible, Michael Buchholz, and Klaus Dietmayer

This paper has been accepted for presentation and publication at the 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), September 24 - 27, 2024, Edmonton, AB, Canada. This is the accepted version of the paper, which has not been fully edited and the layout may differ from the original publication.

Citation information of the original publication:

T. Griebel, A. Scheible, M. Buchholz and K. Dietmayer, ”Towards an Advanced Self-Monitoring Tracking Module: Leveraging Statistical Hypothesis Tests and Subjective Logic Reasoning,” 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), Edmonton, AB, Canada, 2024, pp. 168-175, doi: 10.1109/ITSC58415.2024.10920240.

Towards an Advanced Self-Monitoring Tracking Module: Leveraging Statistical Hypothesis Tests and Subjective Logic Reasoning

Thomas Griebel, Alexander Scheible, Michael Buchholz, and Klaus Dietmayer

***Abstract*— In automated driving systems, monitoring and self-assessment of tracking algorithms is essential. This is especially necessary to meet today's safety and robustness challenges in an automated system. We propose a hybrid approach to develop a self-monitoring module for tracking algorithms. It makes use of well-known statistical hypothesis testing techniques. The results of which are fed into a subjective logic-based reasoning framework to produce robust and reliable self-assessment scores. Hence, we investigate the potential of combining these two approaches for monitoring and self-assessment systems and show the significance of this approach in experimental results.**

I. Introduction

As automated driving algorithms continue to advance towards production development, the challenges to safety and reliability continue to grow. In response, the automotive industry has pushed for compliance with more stringent functional safety standards such as ISO 21448, which addresses Safety of the Intended Functionality (SOTIF) [1]. A key aspect of compliance is the development of Self-Assessment (SA) modules within automated systems, particularly for central tasks such as filtering and tracking algorithms.

Currently, existing filtering and tracking algorithms primarily use consistency tests to monitor and self-assess their algorithms. These consistency tests mostly focus on single criteria, such as the Normalized Innovation Squared (NIS) [2] for Kalman filtering in Single-Object Tracking (SOT). While some extensions of the NIS and other tests for specific criteria have been proposed, these tests can only be performed separately for specific aspects and are not linked. Overall, a comprehensive framework for the development of SA modules is still lacking in the scientific literature.

This work aims to provide a unified SA module and framework for tracking algorithms. The proposed SA module applies classical statistical hypothesis testing approaches [3] for several aspects and assumptions of filter and tracking algorithms. Leveraging the hypothesis testing results, a Subjective Logic (SL) [4] reasoning framework is built on top of that to obtain corresponding SA scores. This is first done for each individual aspect to be tested for the tracking assumption. Then, the individual SA scores are additionally combined and fused in an SL manner to obtain overall SA scores for each sensor and then for the overall tracking algorithm.

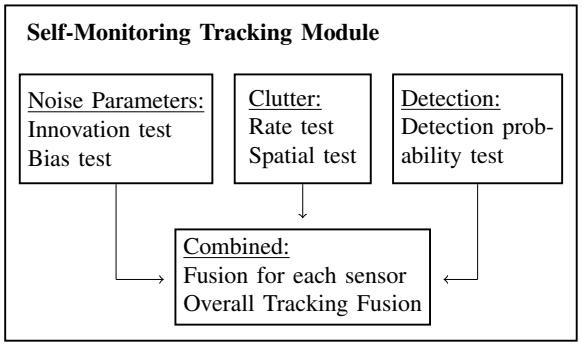

Fig. 1. Our proposed self-monitoring tracking module for single-object tracking in clutter consists of several statistical hypothesis tests for corresponding tracking assumptions, which are then used to build up a subjective logic-based reasoning framework for self-assessment. This results in combined self-assessment scores for each sensor and for the overall tracking system.

In this work, we focus on tracking single objects in clutter, which means that challenges like clutter detections, missed detections, and unknown data associations exist in addition to the classical filtering challenges such as noisy measurements, among others [5]. To build the SA reasoning framework and to be able to unify all the hypothesis test results, SL theory is used, which is a modern extension of probabilistic logic for reasoning under uncertainty. The general concept of the developed SA module is visualized in Fig. 1. With the development of this SA module using statistical hypothesis testing, we are advancing our previous efforts [6]–[9], aiming to contribute to the establishment of a unified and comprehensive SA framework for tracking algorithms.

Summarizing our work in this paper, we propose:

- A unified approach for an SA module for SOT in clutter combining statistical hypothesis testing and SL reasoning framework in Section IV,
- An SA approach that monitors individual and specific tracking aspects and, additionally, overall sensor-specific and general tracking SA scores in Section IV,
- A comprehensive evaluation of the SA module in challenging real-world motivated simulation scenarios in Section V.

Parts of this research have been conducted as part of the EVENTS project and other parts as part of the PoDIUM project, which are both funded by the European Union under grant agreements No. 101069614 and No. 101069547, respectively. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

All authors are with the Institute of Measurement, Control and Microtechnology, Ulm University, Germany, {firstname}.{lastname}@uni-ulm.de

II. Related Work

Classical consistency assessments within Kalman filtering, such as the NIS and the Normalized Estimation Error Squared (NEES), were first introduced by Bar-Shalom et al. [2]. While NEES necessitates access to Ground Truth (GT) data, which is typically unavailable during online applications, the NIS serves as a feasible solution for online assessment. Mahler [10] extended the NIS approach to encompass multi-object scenarios, introducing the Multi-Target NIS (MNIS) and the Multi-Object Generalized NIS (MGNIS), providing divergence detection capabilities without relying on GT data. Reuter et al. [11] further refined this approach for the δ-Generalized Labeled Multi-Bernoulli (GLMB) filter, devising target-dependent (MGNIS$_T$) and clutter-dependent (MGNIS$_C$) variants. To mitigate the clutter dependency of MGNIS$_T$, they proposed the Approximate Multi-Target NIS (AMNIS), aligning its characteristics with the single-target NIS. Additionally, Stübler et al. [12] proposed consistency assessments derived from NIS for feature-based random-set Monte Carlo localization, scrutinizing various components of the measurement model for consistency. Recently, Duník et al. [13] proposed a methodology for assessing tracking reliability by introducing the concept of a reliability index. This index is formulated based on the notion of an ideal Bayesian filter, which offers a GT tracking outcome devoid of any assumptions, approximations, or modeling inaccuracies. Consequently, although theoretically feasible, this index remains unavailable for online estimation. Nonetheless, the field of monitoring and SA of tracking systems remains underexplored, leaving numerous research inquiries unanswered.

On the other hand, more research has been done on classical evaluation measures of Multi-Object Tracking (MOT) systems, for which GT data is required. Some classical MOT evaluation metrics are the optimal subpattern assignment (OSPA) [14] and the generalized OSPA (GOSPA) [15]. Both optimally handle the assignment problem by a parameter c that balances localization and cardinality errors, even considering different cardinalities of multiple object states. In addition, the GOSPA can be divided into three different errors: Localization, missed tracks, and false tracks errors. Using the Hellinger distance, the approach in [16] allows for incorporating track uncertainty into the OSPA metric. To account for errors between the estimated and true set of tracks, the authors of [17] propose the OSPA-on-OSPA (OSPA2) extension of the OSPA. Again, all evaluation techniques require GT data, making them impractical in real-world online applications.

In our previous research, we introduced an SA module for SOT in clutter, incorporating SL theory [7]. This module includes a self-assessing Kalman filter, using SL to characterize statistical uncertainty [6]. Additionally, we expanded this SA framework to encompass linear and nonlinear multi-sensor Kalman filters, enabling the assessment of both individual sensor SAs and overall system performance [8]. In contrast to our current work, the SA techniques developed so far aimed at directly incorporating the SL-based SA approach into the tracking algorithm. This means that the monitoring of the tracking assumption was done directly within the framework of SL theory. The current work, however, aims to leverage all kinds of classical statistical hypothesis testing and then, on top of that, build an SL-based reasoning framework to process and combine the classical testing results.

III. Fundamentals

This section describes the fundamentals of this work. First, the SOT in clutter task is presented mainly based on [5]. Then, classical hypothesis testing is outlined based on [3]. Finally, the basics of SL are introduced based on [4].

A. Single-Object Tracking in Clutter

SOT in clutter represents a specific instance of MOT, where only one object is present and observed in the surroundings. While classical Kalman filtering principles apply, SOT in clutter introduces new challenges, such as handling missed detections, clutter detections, and uncertain data associations alongside noisy measurements and dynamic state estimation. Notably, the primary adaptations from classical Kalman filtering lie in the measurement model and the resulting association, while prediction, update, and object dynamics principles remain typically unchanged. Consequently, our focus centers on the measurement model in the following to address these challenges and the typically made modeling and assumptions for that. Typical algorithms tackling this task are the Nearest Neighbor (NN), the Probabilistic Data Association (PDA), and the Gaussian sum algorithms. In this work, we focus on the NN algorithms. However, the proposed SA methods can also be applied to the other associations algorithms.

At each time step $k \in \mathbb{N}$, we receive from each sensor $s \in \mathbb{S}$ a set of measurements $Z_k^{(s)} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_{m_k}\}$, where $\boldsymbol{z}_i \in \mathbb{R}^m$ for $i = 1, \dots, m_k$. Here, $m \in \mathbb{N}$ denotes the dimension of the measurement space for each individual measurement, and $m_k \in \mathbb{N}_0$ signifies the count of measurements at a time step. The general goal of the filter algorithm is to estimate the object state $\boldsymbol{x}_k \in \mathbb{R}^n$ with the state dimension $n \in \mathbb{N}$ based on the incoming measurements over time. Within Kalman filtering, $\boldsymbol{x}_k$ is modeled by an n-dimensional multivariate Gaussian distribution, characterized by mean $\hat{\boldsymbol{x}}_k \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{P}_k \in \mathbb{R}^{n \times n}$, i.e., $\boldsymbol{x}_k \sim \mathcal{N}(\hat{\boldsymbol{x}}_k, \boldsymbol{P}_k)$. The typically made major assumptions in SOT in clutter algorithm are presented below.

1) Process and Measurement Noise: The noise for the process and measurement model, $\boldsymbol{v}_k \in \mathbb{R}^n$ and $\boldsymbol{w}_k \in \mathbb{R}^m$, respectively, are white, uncorrelated, and Gaussian distributed, i.e., $\boldsymbol{v}_k \sim \mathcal{N}(0, \boldsymbol{Q}_k)$ with $\boldsymbol{Q}_k \in \mathbb{R}^{n \times n}$ and $\boldsymbol{w}_k \sim \mathcal{N}(0, \boldsymbol{R}_k)$ with $\boldsymbol{R}_k \in \mathbb{R}^{m \times m}$.

2) Clutter Detections: Furthermore, clutter detections are modeled by a false alarm process. Here, the number of clutter detections $m_{c_k} \in \mathbb{N}_0$ is Poisson distributed with an expected number $\bar{\lambda}_c \in (0, \infty)$, i.e., $m_{c_k} \sim \text{Poi}(\bar{\lambda}_c)$. All clutter detections are statistically independent and identically distributed with a spatial distribution $\lambda_c(\boldsymbol{z}_k)$. Typically, it is

assumed that $\lambda_c(\boldsymbol{z}_k)$ follows a uniform distribution across the sensor's field of view (FOV) $\mathcal{R}$ with its volume $\text{Vol}(\mathcal{R})$. This means that $\lambda_c(\boldsymbol{z}_k) = \mathcal{U}(\mathcal{R})$ with $\lambda_c(\boldsymbol{z}_k) = \frac{\bar{\lambda}_c}{\text{Vol}(\mathcal{R})}$ for $\boldsymbol{z}_k \in \mathcal{R}$.

3) Missed Detections: The sensors may or may not detect the object at a given time. If the sensor does not detect the object, this is called missed detection. The detection probability $p_D(\boldsymbol{x}_k) \in [0, 1]$ characterizes the probability of a sensor detecting an object at time step k. Consequently, $1-p_D(\boldsymbol{x}_k)$ represents the probability of the sensor failing to detect the object. Thus, the sensor's object detection process conforms to a Bernoulli distribution with probability $p_D(\boldsymbol{x}_k)$, denoted as $\text{Bernoulli}\left(p_D(\boldsymbol{x}_k)\right)$. Note that the detection probability $p_D(\boldsymbol{x}_k)$ is typically assumed to be state-independent and constant such that $p_D(\boldsymbol{x}_k) = p_D$.

B. Hypothesis Testing

Classical hypothesis testing aims to test sampled data against a hypothesis [3]. Thus, first, a hypothesis needs to be formulated based on the questioned goal. The so-called null hypothesis H_0 is, for example, that the sampled data follows a certain distribution. And the alternative or counter hypothesis is that it does not follow the distribution. Then, a significance level $\alpha \in (0, 1)$ needs to be specified. This significance level is the maximum probability that H_0 will be falsely rejected when it is actually true. Typical values for α are 0.05 or 0.01. Then, all of these are used to compute the test statistics. This includes the p-value $p \in [0, 1]$, which denotes the probability, given that H_0 is true, of observing a test statistic at least as extreme as the one from the sampled data. If $p \leq \alpha$, H_0 is rejected and if $p > \alpha$, H_0 is not rejected. This leads to a final decision on the hypothesis testing.

C. Subjective Logic

This section introduces the basics of SL [4], which are needed for our SA reasoning framework. SL is a framework for reasoning under uncertainty and it explicitly accounts for the aspect of statistical uncertainty. The key component of SL is an opinion that represents information about a discrete random variable X in the domain $\mathbb{X}$ with cardinality $|\mathbb{X}| \geq 2$. An SL opinion is defined as

$$\omega_X = (\boldsymbol{b}_X, u_X, \boldsymbol{a}_X), \tag{1}$$

where $\boldsymbol{b}_X$ is the belief mass distribution representing the belief in each event of X, u_X is the uncertainty mass representing the lack of evidence, and $\boldsymbol{a}_X$ is the base rate distribution representing the prior knowledge about X. To form an opinion, the following relations need to be fulfilled:

$$\boldsymbol{b}_X : \mathbb{X} \to [0, 1], \qquad 1 = u_X + \sum_{x \in \mathbb{X}} \boldsymbol{b}_X(x)\,, \tag{2a}$$

$$\boldsymbol{a}_X : \mathbb{X} \to [0, 1], \qquad 1 = \sum_{x \in \mathbb{X}} \boldsymbol{a}_X(x)\,. \tag{2b}$$

When the random variable X has two events in its domain $\mathbb{X} = \{x, \bar{x}\}$, the opinion is called a binomial opinion. This is a special case of the general multinomial opinion case for $|\mathbb{X}| \geq 2$. For binomial opinions, the belief mass $\boldsymbol{b}_X$ can be separated and explicitly expressed as belief $b_x = \boldsymbol{b}_X(x)$ and disbelief $d_x = \boldsymbol{b}_X(\bar{x})$, which yields

$$\omega_X = (b_x, d_x, u_X, \boldsymbol{a}_X). \tag{3}$$

Opinions can be mapped to the classical probability space using the projected probability, i.e.,

$$\boldsymbol{P}_X(x) = \boldsymbol{b}_X(x) + \boldsymbol{a}_X(x) u_X, \quad \forall x \in \mathbb{X}. \tag{4}$$

The projected probability $\boldsymbol{P}_X$ equals the expected outcome of the opinion in the probability space using Dirichlet distributions [4].

One of the advantages of SL is its powerful fusion framework. Here, multiple opinions $\omega_X^{S_1}, \ldots, \omega_X^{S_N}$ from different sources $S_1, \ldots S_N$ about the same random variable $X \in \mathbb{X}$ can be merged together to form a comprehensive fused statement in terms of an opinion $\omega_X^{\mathbb{S}}$. Within SL, a variety of fusion operators exist. The choice of which fusion operator is suitable depends on the considered applications within its situation and their underlying assumptions. Common fusion operators are, for example, the cumulative belief fusion (CBF) and the averaging belief fusion (ABF) [4]. The CBF operator is denoted as $\omega_X^{\diamond(\mathbb{S})}$ and abbreviated as '$\oplus$', and is given by [4]

$$\omega_X^{\diamond(\mathbb{S})} = \bigoplus_{S \in \mathbb{S}} \left(\omega_X^S\right) = \omega_X^{S_1} \oplus \ldots \oplus \omega_X^{S_N}. \tag{5}$$

It is particularly suitable when the information from different sources, which are merged, is independent of each other. This also means that more evidence should decrease the uncertainty in this case. On the other hand, the ABF operator is denoted as $\omega_X^{\underline{\diamond}(\mathbb{S})}$ and abbreviated as '$\underline{\oplus}$' using [4]

$$\omega^{\underline{\diamond}(\mathbb{S})} = \underset{S \in \mathbb{S}}{\underline{\oplus}} \left(\omega_X^S\right). \tag{6}$$

ABF is suitable when the information of sources is not independent of each other. This means that more information does not necessarily decrease the uncertainty. For more information about SL, please refer to [4].

IV. Self-Monitoring Tracking Module

This section presents the proposed SA module for monitoring various assumptions and aspects of the NN tracking algorithm for the task of SOT in clutter. First, several statistical hypothesis tests are applied to the assumptions stated in Section III-A, such as the innovation, bias, clutter rate, spatial clutter distribution, and detection probability. Based on these hypothesis tests, an SL opinion for each result is obtained. These SL opinions can then be fused to compute a combined SA score. If multiple sensors are available within the perception system, this SA score can be computed individually for each sensor. The SA scores for all sensors can be fused again in order to obtain an overall tracking SA score. This SA module is visualized in Fig. 1 and outlined in the following.

A. Innovation Test

The innovation test is based on the NIS consistency test for Kalman filtering, which focuses on testing the assumptions about the process and the measurement noise described in Section III-A.1. The NIS is computed by [18]

$$\varepsilon_{\boldsymbol{\gamma}_k} = \boldsymbol{\gamma}_k^T \boldsymbol{S}_k^{-1} \boldsymbol{\gamma}_k. \tag{7}$$

Here, $\boldsymbol{\gamma}_k = \boldsymbol{z}_k - \hat{\boldsymbol{z}}_{k|k-1}$ is the innovation with the measurement prediction with mean $\hat{\boldsymbol{z}}_{k|k-1} \in \mathbb{R}^m$ and its innovation covariance $\boldsymbol{S}_k \in \mathbb{R}^{m \times m}$. Then, if the models are linear and the assumptions about the process and measurements noise are met, $\varepsilon_{\boldsymbol{\gamma}_k}$ is χ^2 distributed with m degrees of freedom.

For hypothesis testing, the null hypothesis H_0 is defined as "The innovation $\boldsymbol{\gamma}_k$ is consistent with the innovation covariance $\boldsymbol{S}_k$", which also includes the assumptions stated in Section III-A.1. H_0 is accepted if $\varepsilon_{\boldsymbol{\gamma}_k} \in [r_1, r_2]$, with the acceptance interval $[r_1, r_2]$ calculated such that the probability accepting H_0 is $1-\alpha$, i.e., $P\big[\varepsilon_{\boldsymbol{\gamma}_k} \in [r_1, r_2] | H_0\big] = 1-\alpha$. This means the acceptance interval $[r_1, r_2]$ is calculated using the inverse cumulative distribution function F^{-1} of the χ^2 distribution, i.e.,

$$r_1 = F^{-1}\left(\frac{\alpha}{2}, m\right), \quad r_2 = F^{-1}\left(1 - \frac{\alpha}{2}, m\right). \tag{8}$$

Because the χ^2 distribution is asymmetric for small degrees of freedom, a one-sided hypothesis test is often used here, which yields a lower bound of $r_1 = 0$.

The results are used to build up the binomial opinion $\omega_{X_{\text{inno}}} = (b_x, d_x, u_X, \boldsymbol{a}_X)$ of the random variable $X_{\text{inno}} \in \mathbb{X}_{\text{inno}} = \{x_{H_0}, \bar{x}_{H_0}\}$. Here, the event x_{H_0} corresponds to the acceptance of H_0 and $\bar{x}_{H_0}$ to the rejection of H_0. Then, at each time step k, one result of the innovation hypothesis test is obtained, which results in evidence of either $x_{H_0} = 1$ or $\bar{x}_{H_0} = 1$. Note that the opinion $\omega_{X_{\text{inno}}}$ is initialized with the uncertainty $u_X = 1$ to account for the fact that no evidence has been collected at the beginning, and the base rates $\boldsymbol{a}_X(x_{H_0}) = 1 - \alpha$ and $\boldsymbol{a}_X(\bar{x}_{H_0}) = \alpha$ using the given significance level α of the hypothesis test. Then, using the bijective mapping from SL [4], which maps the collected evidence to an opinion, $\omega^k_{X_{\text{inno}}}$ is obtained. These opinions are fused together over a given sliding window of length $n_{\text{s}} \in \mathbb{N}$ using the CBF operator (5) to accumulate evidence over time. This yields a sliding window innovation opinion $\omega_{X_{\text{inno}}}$. Note that the time index k and the corresponding indices for the sliding window consideration are ignored in the following for reasons of clarity. Calculating the projected probability of $\omega_{X_{\text{inno}}}$ with (4), the innovation test SA score $P_{X_{\text{inno}}}(x_{H_0}) \in [0, 1]$ is obtained. A high score near 1 means that the innovation test states that the tested assumptions in terms of H_0 are fulfilled, and the SA reports that the algorithm is working as expected. In contrast, a low score near 0 means that the assumptions to be tested in terms of H_0 are likely to be not fulfilled, such that the SA reports assumption violations, which can lead to unreliable tracking estimates.

B. Bias Test

The bias test is similar to the innovation test. In fact, the NIS also implicitly tests for a measurement bias. This means the bias test is a targeted version of the innovation test towards a bias in the assumptions of Section III-A.1. Therefore, the normalized mean innovation is calculated as

$$\varepsilon_{\boldsymbol{\mu}_{k,j}} = \frac{\boldsymbol{\gamma}_{k,j}}{\sqrt{\boldsymbol{S}_{k,(j,j)}}} \tag{9}$$

for all measurement components $j = 1, \ldots, m$. Note that $\boldsymbol{S}_{k,(j,j)}$ is the scalar component at row j and column j of the innovation covariance $\boldsymbol{S}_k$. If the innovation is bias-free, then $\varepsilon_{\boldsymbol{\mu}_{k,j}} \sim \mathcal{N}(0, 1)$ for all $j = 1, \ldots, m$.

Then, for the hypothesis testing, the null hypothesis H_0 states: "The measurements, and thus the innovations, are bias-free". H_0 is accepted if $\varepsilon_{\boldsymbol{\mu}_{k,j}} \in [-r, r]$, $\forall j$, due to the symmetry of the standard normal distribution. The acceptance interval $[-r, r]$ is similarly calculated as before with the significance level α for the standard normal distribution. The hypothesis testing results are used to create the binomial opinion $\omega_{X_{\text{bias}}}$ with $X_{\text{bias}} \in \mathbb{X}_{\text{bias}} = \{x_{H_0}, \bar{x}_{H_0}\}$. Following the same procedure as before, using a sliding window approach for $\omega_{X_{\text{bias}}}$ with the CBF operator and n_{s} time steps, the bias test SA score is obtained by the projected probability $P_{X_{\text{bias}}}(x_{H_0}) \in [0, 1]$. Note that the sliding window length n_{s} can be chosen differently than before, but for the sake of simplicity, it is denoted as the same.

C. Clutter Rate Test

The clutter rate test focuses on the fulfillment of the assumption that the clutter rate follows a Poisson distribution with an expected number $\bar{\lambda}_c$, i.e., $m_{c_k} \sim \text{Poi}(\bar{\lambda}_c)$, from Section III-A.2. This yields the null hypothesis H_0: "The clutter rate m_{c_k} is Poisson distributed with expected value $\bar{\lambda}_c$". H_0 is accepted if $m_{c_k} \in [r_1, r_2]$, with the acceptance interval $[r_1, r_2]$ of the Poisson distribution computed such that the probability accepting H_0 is $1 - \alpha$ as before. This means the acceptance interval $[r_1, r_2]$ is calculated using the inverse of the cumulative distribution function (or percent point function) F^{-1} of the Poisson distribution similar to (8). Then, the evidence for the number of clutter measurements is obtained by

$$m_{c_k} = \begin{cases} 0, & \text{for } m_k = 0, \\ m_k - p_D, & \text{for } m_k > 0. \end{cases} \tag{10}$$

Using the obtained evidence, the hypothesis testing for H_0 is performed to build up the binomial opinion $\omega_{X_{\text{clut-rate}}}$ with $X_{\text{clut-rate}} \in \mathbb{X}_{\text{clut-rate}} = \{x_{H_0}, \bar{x}_{H_0}\}$. Again, following the same procedure as before, using a sliding window approach for $\omega_{X_{\text{clut-rate}}}$ with the CBF operator and n_{s} time steps, the clutter rate test SA score is obtained by the projected probability $P_{X_{\text{clut-rate}}}(x_{H_0}) \in [0, 1]$.

D. Spatial Clutter Distribution Test

The spatial clutter distribution test monitors the fulfillment of the assumption that clutter detections are uniformly spatially distributed, i.e., that $\lambda_c(\boldsymbol{z}_k) = \mathcal{U}(\mathcal{R})$, with $\lambda_c(\boldsymbol{z}_k) =$

$\frac{\bar{\lambda}_c}{\text{Vol}(\mathcal{R})}$ for $\boldsymbol{z}_k \in \mathcal{R}$ from Section III-A.2. This test is performed using a Kolmogorov–Smirnov Goodness-of-Fit test [19]. This test compares the underlying distribution of obtained samples against the assumed distribution. The null hypothesis H_0 is defined as "The clutter is uniformly spatially distributed", as stated above. The samples for this test are obtained by considering all incoming measurements $Z_k^{(s)} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{m_k}\}$ in sensor coordinates for sensor s. First, the NN-associated measurement is removed from the measurement set as it is assumed to be the object-originated one. Note that the NN-associated measurement is the measurement selected by the NN association algorithm as the object-originated measurement, associated with the object track, and then updated accordingly. Then, the measurements are scaled and normalized using the sensor FOV to be mapped on the range $[0, 1]$. Finally, the scaled measurement set is input to the Kolmogorov–Smirnov Goodness-of-Fit test and tested to a uniform distribution on $[0, 1]$. This test outputs a p-value p as described in Section III-B. The p-value p is compared to the chosen significance level α, and, in this way, evidence is collected for supporting or rejecting H_0. Using this, the binomial opinion $\omega_{X_{\text{clut-spatial}}}$ with $X_{\text{clut-spatial}} \in \mathbb{X}_{\text{clut-spatial}} = \{x_{H_0}, \bar{x}_{H_0}\}$ is created. Following the same procedure as for the other tests, using the sliding window approach for $\omega_{X_{\text{clut-spatial}}}$ with the CBF operator and n_s time steps, the clutter spatial distribution test SA score is obtained by the projected probability $P_{X_{\text{clut-spatial}}}(x_{H_0}) \in [0, 1]$.

E. Detection Probability Test

The detection probability test focuses on the monitoring of the assumption that the sensor's object detection process follows a Bernoulli distribution with constant probability p_D, i.e., Bernoulli (p_D). Here, a binomial test [19] is performed, which is an exact test for a binomial distribution. This means that the evidence of a performed detection (one measurement is NN-associated at a time step) is accumulated over time here using a sliding window of n_s time steps to perform a more statistically significant test. Thus, the null hypothesis H_0 is "The detection process over time follows a binomial distribution with probability p_D, i.e., Binomial$(n_{\text{dets}}, n_\text{s}, p_D)$ with the number of NN-associated detections n_{dets} in the time window n_s", which corresponds to a sequence of Bernoulli experiments. Then, using n_{dets}, the binomial test can be performed. The test output result of the p-value p is compared to the chosen significance level α, and, in this way, evidence is collected for supporting or rejecting H_0. Based on this, a binomial opinion $\omega_{X_{\text{det}}}$ with $X_{\text{det}} \in \mathbb{X}_{\text{det}} = \{x_{H_0}, \bar{x}_{H_0}\}$ is created. Then, following the same procedure as for the other tests, the detection probability test SA score is obtained by the projected probability $P_{X_{\text{det}}}(x_{H_0}) \in [0, 1]$.

F. Combined Hypotheses Tests for Each Sensor

All these statistical tests can be performed for one sensor to test all the assumptions' fulfillment. Using the obtained SL opinions for each test, a combined SA score can be obtained by fusing the opinions in the SL reasoning framework. This is performed by using the CBF operator (5) to accumulate evidence and to obtain $\omega_{X_{\text{comb}}}$ with the random variable X_{comb} combining all events for accepting H_0, the union of all individual events x_{H_0}, and its counter-events for rejecting H_0, the union of all individual events $\bar{x}_{H_0}$, in the domain $\mathbb{X}_{\text{comb}} = \{x_{H_0}, \bar{x}_{H_0}\}$. This yields the combined opinion

$$\omega_{X_{\text{comb}}} = \omega_{X_{\text{inno}}} \oplus \omega_{X_{\text{bias}}} \oplus \omega_{X_{\text{clut-rate}}} \oplus \omega_{X_{\text{clut-spatial}}} \oplus \omega_{X_{\text{det}}}. \tag{11}$$

Then, calculating the projected probability of $\omega_{X_{\text{comb}}}$ with (4), the combined SA score for one sensor $P_{X_{\text{comb}}}(x_{H_0}) \in [0, 1]$ is obtained.

G. Overall Tracking Monitoring

When multiple sensors are involved in the tracking system, the combined SA score can be calculated for each sensor $s \in \mathbb{S}$. This results in combined opinions $\omega_{X_{\text{comb}}}^{(s)}$ for all sensors. These combined opinions can then be fused in the SL reasoning framework to obtain one overall SA score for the whole tracking algorithm. This leads to the opinion $\omega_{X_{\text{overall}}}$ with $X_{\text{overall}} \in \mathbb{X}_{\text{overall}} = \{x_{H_0}, \bar{x}_{H_0}\}$ by using the ABF operator (6) to average the collected evidence. The ABF operator is used here to average the combined opinions of all sensors in order to obtain an overall averaged tracking SA score. The resulting opinion is calculated by

$$\omega_{X_{\text{overall}}} = \underset{s \in \mathbb{S}}{\oplus} \left(\omega_{X_{\text{comb}}}^{(s)} \right). \tag{12}$$

Again, the overall tracking SA score can be calculated by projection, yielding $P_{X_{\text{overall}}}(x_{H_0}) \in [0, 1]$.

V. Experiments

In this section, we evaluate our proposed self-monitoring module in SOT in clutter using the NN association algorithm. First, we consider a scenario with multiple disturbances in different aspects and tracking assumptions. Second, a scenario is set up to simulate disturbances caused by adverse weather conditions and mirroring effects in an urban environment.

In our simulations, we analyze a multi-sensor system comprising three sensors that measure the objects in two dimensions $(x, y) \in \mathbb{R}^2$. To evaluate our SA module approach, we initially assume that the assumptions regarding all three sensors in the NN tracking algorithm are identical to the GT-modeled assumptions for the simulation. This initial assumption allows our SA to affirm that the system operates as expected. Additionally, we employ the nearly constant velocity (CV) model [18], resulting in a linear scenario. All results shown are averaged values from 100 Monte Carlo runs. For our proposed self-monitoring tracking module, we choose the significance level $\alpha = 0.05$ and the time window length $n_\text{s} = 35$ for all hypothesis tests.

A. Disturbance Scenario

First, a scenario with multiple disturbances is considered. Here, the focus and most of the disturbances are in GT-modeled assumptions of Sensor 1. The overview of the consecutively injected disturbances of the scenario is given in Table I. The arrow pointing up '↑' and the arrow pointing

TABLE I

OVERVIEW OF THE DISTURBANCES IN THE FIRST DISTURBANCE SCENARIO.

Time steps / Sensors	100 – 200	300 – 400	500 – 600	700 – 800	900 – 1000	1100 – 1200	1300 – 1400
Sensor 1	↑ meas. noise	↑ clutter rate	Δ spatial clutter dist.	↓ det. prob.	+ meas. bias		
Sensor 2						+ meas. bias	
Sensor 3							+ meas. bias

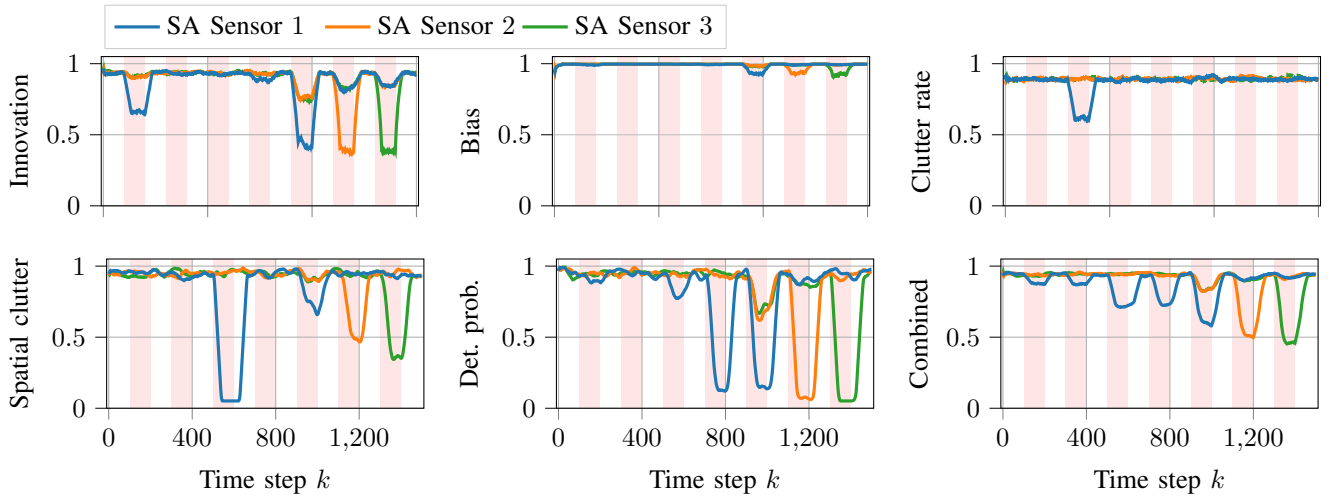

Fig. 2. The self-monitoring tracking module results, consisting of five SA tests and the combined SA scores for all three sensors, are shown. Table I shows the corresponding disturbances of the scenario. Red shadows indicate these disturbances, which are monitored by the SA module.

down '↓' mean that the related assumption is violated by increasing and decreasing the corresponding GT simulated parameter, respectively. The 'Δ' means that the spatial clutter distribution is changed from the assumed uniform distribution to Gaussian. Moreover, the '+' means a measurement bias is added to the measurements. These disturbances are injected in the given corresponding time interval. Outside of these disturbance intervals, all assumptions are set back to the original ones, thus satisfying the filter assumptions. Note that the filter assumes the same initial assumption throughout the scenario, which leads to violations of the assumptions in the corresponding disturbance intervals.

The results of the proposed self-monitoring tracking module for all three sensors are presented in Fig. 2. The results show that the SA module is able to monitor its assumption and detect the corresponding violations. Especially the measurement bias violations in all sensors are strong disturbances that lead to multiple SA test violation reports indicated by the low and decreasing SA scores. Note that the bias violation is so significant that there are multiple non-associated measurements during these time intervals, resulting in missed detections. This is indicated by the low detection probability SA scores. However, this has only a small effect on the bias SA score itself because if there is a missed detection, no evidence is generated for the bias SA. In Fig. 3, the comparison results of a time-averaged variant of the NIS, the overall SA tracking score from Section IV-G, and the positional root mean squared error (RMSE) as a GT-based evaluation measure are shown. It can also be seen here that the measurement bias violations are significantly visible in the time-average NIS. Moreover, the RMSE also shows that the bias violations in all three sensors lead to the biggest errors in the scenario. Due to the averaging fusion in the overall SA, some declines in individual SA scores are averaged out. Note that not all disturbances and assumption violations lead to some performance degradation in the RMSE metric. For this, a sensitivity analysis between assumption violations and performance degradation in evaluation metrics is important in future works towards a self-assessing performance degradation estimation.

B. Disturbances in Urban Environments

The next simulation scenario is motivated by disturbances caused by adverse weather conditions and mirroring effects in urban environments. The case of adverse weather in an urban environment can lead to various disturbances in the sensors typically used in automated driving systems. Namely, for lidar sensors, these conditions can lead to an increased noise of the detections, a decreased object detection rate, and an increased clutter rate. Radar sensors, in contrast, are mostly unaffected by these conditions. However, in some cases, an increase in the number of radar clutter detections can be caused by urban environments. Moreover, camera sensors may be similarly affected to lidar sensors, except for the increased clutter rate. In addition, reflective conditions in an urban environment, e.g., large glass walls, usually tend to increase the number of clutter detections by all three sensors

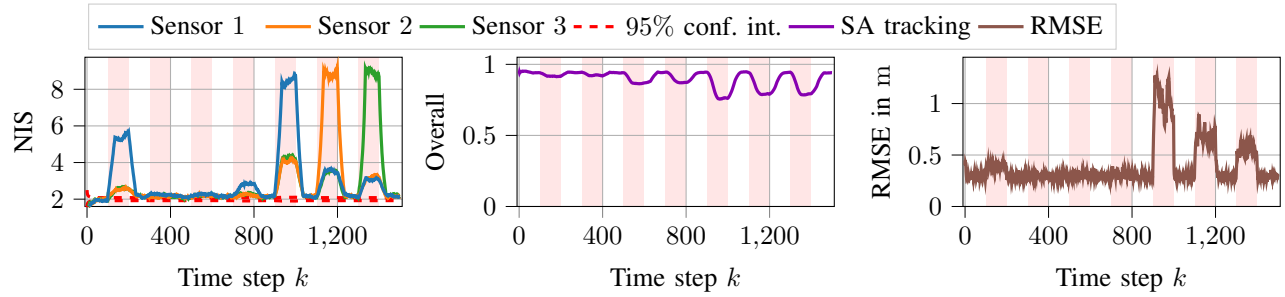

Fig. 3. The results of the time-averaged NIS as a comparison measure, our obtained overall SA tracking score based on all sensors' monitoring modules, and the GT evaluation RMSE measure are shown. The disturbances of the scenario are summarized in Table I and highlighted by the red shadows.

TABLE II

OVERVIEW OF THE DISTURBANCES IN AN URBAN ENVIRONMENT SCENARIO WITH ADVERSE WEATHER CONDITIONS AND MIRRORING EFFECTS.

Time steps / **Sensors**	$100-200$ (adverse weather)	$300-400$ (mirroring effects)
Sensor 1 - Lidar	↑ meas. noise & ↓ det. prob. & ↑ clutter rate	↑ clutter rate
Sensor 2 - Radar	↑ clutter rate	↑ clutter rate
Sensor 3 - Camera	↑ meas. noise & ↓ det. prob.	↑ clutter rate

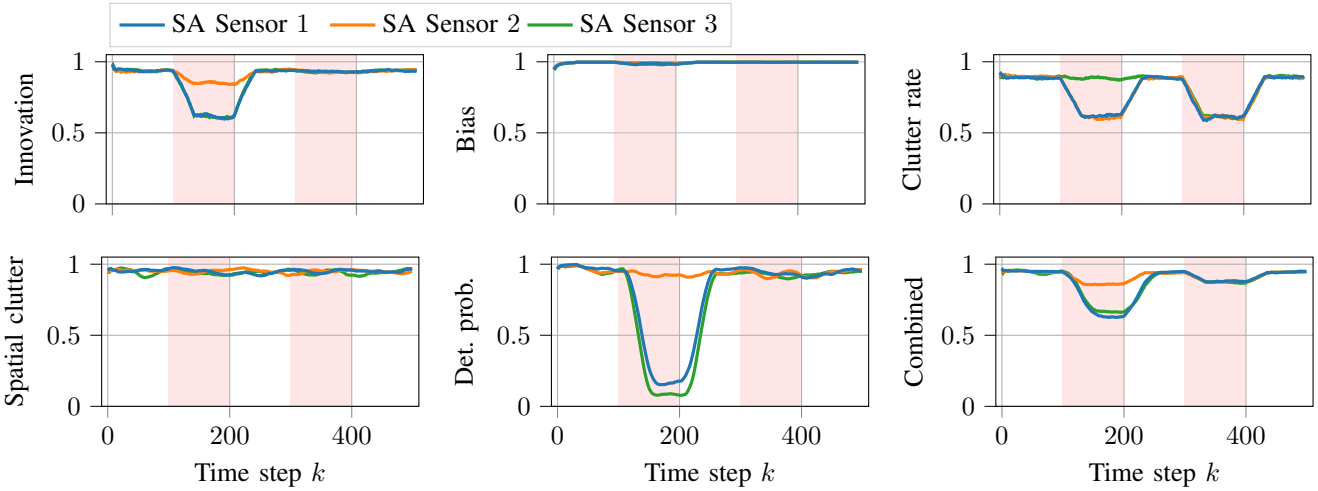

Fig. 4. The self-monitoring tracking module results, consisting of five SA tests and the combined SA scores for all three sensors, are shown. Table II shows the corresponding disturbances of the urban environment scenario. Red shadows indicate these disturbances, which are monitored by the SA module.

due to the mirroring effect. This scenario is summarized in Table II. First, from time step $100-200$, the disturbances of adverse weather are simulated, and then from time step $300-400$, the disturbances of mirroring effects by many reflective surfaces are simulated.

The results of the proposed self-monitoring tracking module for all three sensors in the urban environment scenario with multiple simultaneous disturbances are shown in Fig. 4. Here, it can be seen that even with multiple simultaneous disturbances, the SA module is able to correctly monitor these violations in the corresponding aspects. In contrast, the time-averaged NIS measure shown for comparison in Fig. 5 gives only a single score with its confidence interval, where the source of the disturbance causing the increased NIS value is not identifiable. Moreover, the RMSE signifies that the error is mostly increased in the adverse weather disturbance interval, which also leads to a larger decrease in the overall SA score. From this, it can be concluded that the self-monitoring tracking module is capable of correctly monitoring not only violations of assumptions in the corresponding categories in a timely and consistent manner but also multiple failures at the same time. Still, it is also able to fuse the individual SA scores in an SL reasoning manner to obtain combined SA scores for each sensor and an overall SA score for the whole tracking algorithm.

VI. CONCLUSION

This paper proposed a hybrid approach for the development of a self-monitoring tracking module. In fact, statistical hypothesis tests are applied, and, in addition, the test outputs are fed into an SL reasoning framework to obtain the SA results. Following this approach, many more insights can

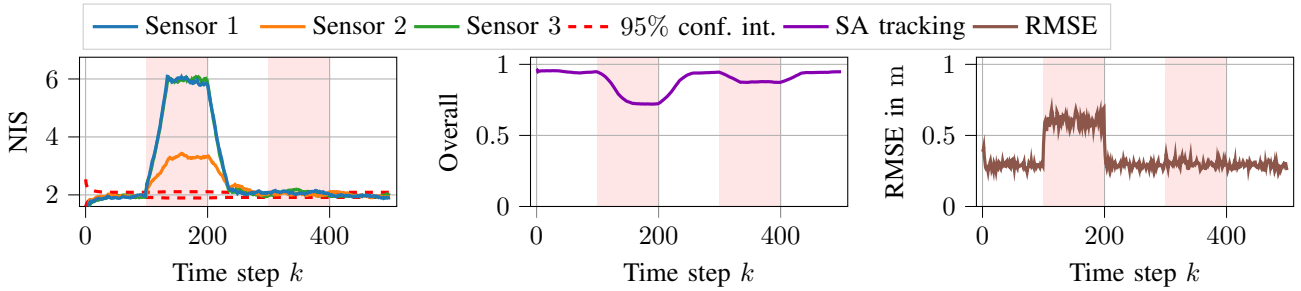

Fig. 5. The results of the time-averaged NIS as a comparison measure, our obtained overall SA tracking score based on all sensors' monitoring modules, and the GT evaluation RMSE measure are shown. The disturbances of the urban environment scenario are summarized in Table II and highlighted by the red shadows.

be obtained than from the classical comparison method of the time-averaged NIS since the individual and the overall results can be used for further decision-making on the state of health of the perception system. In challenging simulation experiments, our self-monitoring tracking module showed good results in self-assessing the disturbances, which have been chosen to be close to real-world effects in automated vehicles. However, not all disturbances and assumption violations lead to the same performance degradation in the evaluation metrics. Hence, an important future research field is to connect the obtained SA measures with the GT evaluation metrics, which evaluate the actual performance degradation of the filter estimates. Further future work will include real-world testing on real-world data from automated vehicles as well as the extension of the method towards MOT.

References

[1] "ISO/PAS 21448: Road vehicles - Safety of the intended functionality," *International Organization for Standardization*, 2019.

[2] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*. Academic Press, New York, 1988.

[3] D. E. Knuth, *The art of computer programming*. Pearson Education, 1997, vol. 3.

[4] A. Jøsang, *Subjective Logic: A formalism for reasoning under uncertainty*. Springer International Publishing, 2016.

[5] S. Challa, M. R. Morelande, D. Mušicki, and R. J. Evans, *Fundamentals of Object Tracking*. Cambridge University Press, 2011.

[6] T. Griebel, J. Müller, M. Buchholz, and K. Dietmayer, "Kalman filter meets subjective logic: A self-assessing Kalman filter using subjective logic," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, 2020, pp. 1–8.

[7] T. Griebel, J. Müller, P. Geisler, C. Hermann, M. Herrmann, M. Buchholz, and K. Dietmayer, "Self-assessment for single-object tracking in clutter using subjective logic," in *2022 25th International Conference on Information Fusion (FUSION)*. IEEE, 2022, pp. 1–8.

[8] T. Griebel, J. Heinzler, M. Buchholz, and K. Dietmayer, "Online performance assessment of multi-sensor Kalman filters based on subjective logic," in *2023 26th International Conference on Information Fusion (FUSION)*. IEEE, 2023, pp. 1–8.

[9] T. Griebel, N. Dehler, A. Scheible, M. Buchholz, and K. Dietmayer, "Self-assessment for multi-object tracking based on subjective logic," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 1750–1757.

[10] R. Mahler, "Divergence detectors for multitarget tracking algorithms," in *Signal Processing, Sensor Fusion, and Target Recognition XXII*, vol. 8745. SPIE, 2013.

[11] S. Reuter, B.-T. Vo, B. Wilking, D. Meissner, and K. Dietmayer, "Divergence detectors for the δ-generalized labeled multi-Bernoulli filter," in *2013 Workshop on Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. IEEE, 2013, pp. 1–6.

[12] M. Stübler, S. Reuter, and K. Dietmayer, "Consistency of feature-based random-set Monte-Carlo localization," in *2017 European Conference on Mobile Robots (ECMR)*, 2017, pp. 1–6.

[13] J. Duník, O. Straka, and B. Noack, "Classification of uncertainty sources for reliable Bayesian estimation," in *2023 IEEE Symposium Sensor Data Fusion and International Conference on Multisensor Fusion and Integration (SDF-MFI)*, 2023, pp. 1–8.

[14] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, 2008.

[15] A. S. Rahmathullah, Á. F. García-Fernández, and L. Svensson, "Generalized optimal sub-pattern assignment metric," in *2017 20th International Conference on Information Fusion (Fusion)*. IEEE, 2017, pp. 1–8.

[16] S. Nagappa, D. E. Clark, and R. Mahler, "Incorporating track uncertainty into the ospa metric," in *14th International Conference on Information Fusion*. IEEE, 2011, pp. 1–8.

[17] M. Beard, B. T. Vo, and B.-N. Vo, "Ospa (2): Using the ospa metric to evaluate multi-target tracking performance," in *2017 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2017, pp. 86–91.

[18] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2001.

[19] W. Daniel, *Applied Nonparametric Statistics*, ser. Duxbury advanced series in statistics and decision sciences. PWS-KENT Pub., 1990.