14th International Conference on Information Fusion
Chicago, Illinois, USA, July 5-8, 2011

Performance Measures for Multiple Target Tracking Problems

A. A. Gorji, R. Tharmarasa, and T. Kirubarajan
Department of Electrical Engineering
McMaster University, Hamilton, Canada

***Abstract*—Performance evaluation is one of the most important steps in any target tracking problem. The objective of this paper is to present a brief review of different approaches available for the performance analysis of multiple target tracking algorithms. Metrics are first classified into *sensor-dependent* and *tracker-dependent* ones. Then, the tracker-dependent measures are discussed after classifying into two groups named *algorithm-free* and *algorithm dependent* measures. For the classification purposes, three different categories of algorithm-free metrics are described. Finally, to demonstrate the application of the metrics in evaluating the performance of different tracking algorithms, a challenging scenario is considered.**

Keywords: Performance evaluation, track to truth association, metric classification.

I. Introduction

Performance evaluation is one of the most important steps in any target tracking problem where the main goal is to compare the performance of different tracking algorithms in a general Multiple Target Tracking (MTT) problem. The variety of available tracking approaches, the recent advances in estimation methods and the necessity to evaluate the accuracy and efficiency in real scenarios make it necessary to study potential Measures of Performance (MOP) for tracking algorithms.

Several categories of MOPs are identified in the literature where metrics have been usually defined according to the type of the performance evaluated by the metric, or the availability of input data. Given the source of input data, metrics are divided into two general sub-classes. MOPs can be *sensor* or *tracker* related where the first class is concerned with the characteristics of the measurement sensors and the second class utilizes tracker results in order to evaluate the accuracy and quality of estimates. Although other classifications may be found in the literature [3] [8], all classes fall under the sensor or tacker related categories.

Even though performance metrics have been defined for individual trackers and applications such as Multiple Hypothesis Tracking (MHT) [4], Track-Before-Detect (TBD) [16], Interactive-Multiple-Model (IMM) [12] [18] [21], Integrated-Data-Association-Filter (IPDA) [22], and Joint Probability Data Association Filter (JPDAF) [20], the *algorithm-free* metrics that can be applied to every tracker are of more interest. MOPs have been defined in terms of cardinality, time and accuracy measures in [15] for a general MTT problem. A more comprehensive work was presented in [13] where the performance of estimation algorithms was evaluated by representing a number of accuracy metrics. In [10] [7], another class of performance metrics was introduced for the multisatic working group (MSWG) where the measures were classified according to the type of input data (measurements or tracking results). The metrics from the MSWG can be also considered as a subclass of the three categories mentioned above.

In this review paper, a general classification of MOPs is represented so that it is applicable to general MTT problems. Measures are divided into track and sensor-related metrics based on their dependence on the tracking data. The focus is on the track-related metrics where measures are classified into algorithm-based and algorithm-free metrics. After a brief description of algorithm-based approaches, algorithm-free metrics which can be applied to every tracking algorithm are presented. A comprehensive review of available algorithm-free metrics is made with a solid classification of metrics into time, cardinality, and accuracy categories.

The rest of this paper is organized as follows. Section II deals with the general MTT problem. Performance metric classification is done in Section III. The main part of this paper is given in Section IV where different types of algorithm-free metrics are described. Section V presents statistical tests for the performance evaluation in the absence of truth. Simulation results are provided in Section VI on a challenging MTT scenario being used as the benchmark. Finally, Section VII concludes the paper.

II. Multiple Target Tracking

The following model can be used to characterize the motion of each target as well as the received measurements [1]

$$\begin{aligned}\mathbf{x}(k+1) &= F\mathbf{x}(k) + \Gamma\mathbf{v}(k+1)\\ \mathbf{y}(k+1) &= \mathbf{h}(\mathbf{x}(k+1)) + \mathbf{w}(k+1)\end{aligned} \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ represent target state and the measurement, respectively, $\mathbf{u}$ denotes the external input, and $\mathbf{v}$ and $\mathbf{w}$ are the additive state and measurement noises, respectively. In above equation, F and Γ are system matrices characterizing the movement of the target, and both additive noises are Gaussian with zero mean and covariance matrices Σ^v and Σ^w, respectively. Also, $\mathbf{h}$ describes the nonlinear sensor model. In MTT, the final goal is to estimate states and number of targets in the surveillance region. Now, consider a general MTT scenario as defined by the above equation in K steps where $k = 1, 2, ..., K$ denotes the index of each scan. The time stamp for each step is demonstrated by t_k where the time stamp may be different for each step. The number of targets at

each step is assumed to be $L(k)$. Assume that truths are known where performance evaluation for the case of unknown truths is also discussed in the paper. The number of measurements at the k-th step is also shown by $m(k)$. In addition, $\mathcal{M}(k)$ denotes the set of all measurements at the k-th step. Now, define $N(k)$ as the number of tracks at the k-th step. Each track is denoted by $\mathcal{T}_n(k)$ best summarizing the information content of measurements in the corresponding step [11]. The characterization of $\mathcal{T}$ depends on the type of the tracker. For example, for an arbitrary tracking algorithm such as Joint Probability Data Association Filter (JPDAF) [20] working with Extended Kalman Filter (EKF), $\mathcal{T}_n(k)$ is defined as

$$\mathcal{T}_n(k) = \{t_k, \mathbf{x}_n(k|k), P_n(k|k)\} \tag{2}$$

where $\mathbf{x}_n(k|k), P_n(k|k)$ represent the mean and covariance of estimation for the n-th track, respectively. The same definition can be given for other trackers. If other filtering methods such as particle filters are used, particles and associated weights replace the estimated mean and covariance matrix. As an example, for a Monte Carlo JPDAF, the following definition is obtained

$$\mathcal{T}_n(k) = \left\{t_k, \mathbf{x}^i(k), \hat{w}_i(k)\right\} \tag{3}$$

with $i = 1, ..., N_p$ and N_p as the number of particles. The set of all tracks at the k-th step is defined as

$$\mathcal{T}(k) = \left\{\mathcal{T}_1(k), \mathcal{T}_2(k), ..., \mathcal{T}_{N(k)}(k)\right\} \tag{4}$$

Additionally, $\mathcal{T}^*(K^*)$ is defined as the collection of all tracks in the interval $k \in [k_{min}, K^*]$. It is worth noting that, $\mathcal{T}(k)$s may have different sizes because the number of tracks at each time step may change by time. The number of estimated tracks ($N(k)$) is not also necessarily equal to that of truths ($L(k)$) [1]. The same set can be defined for truths holding the states of targets. The following expression is defined for $\mathcal{L}_l(k)$ at the k-th step:

$$\mathcal{L}_l(k) = \left\{t_k^l, \mathbf{x}_l(k)\right\} \tag{5}$$

where t_k^l is the l-th target appearance time. Note that the target's appearance time is not necessarily equal to t_k, which is represented as the simulation time. The simulation sample time is usually larger than the target's sample time due to the sensors' limitations. MOPs might be defined according to the target appearance time or the simulation time based on the application and availability of truths and measurements. For example, when truths and tracks are both available, metrics are found during the time that both sources of information are available, which is usually the simulation time.

III. Performance Measures

Different classes of performance metrics can be defined based on the availability of inputs to the performance evaluator. The following classes of metrics can be defined:

[1]The reason is the presence of false tracks. Indeed, every tracking system may estimate some tracks not assigned to actual targets. In some other cases, the tracker may miss the truth and, consequently, the number of estimates and truths may be completely different.

- Sensor-related measures:
  Measurements and truths are used to construct metrics for the MTT problem. The metric function can be written as

$$\mathbf{p}_{SR} = \mathbf{g}(\mathcal{M}, \mathcal{L}) \tag{6}$$

  There are also some situations that the best achievable performance of a tracking algorithm is to be found. In this situation, the Posterior Cramer Rao Lower Bound (PCRLB) [1] of tracking may be derived in order to find a minimum bound of tracking estimation, which is independent of the tracking algorithm.
- Track-related measures:
  Unlike the sensor-related measures, the tracker is also taken into consideration for metric construction. The following sub-categories are considered for the track related measures:
  - Algorithm-based:
    These measures are developed for individual types of trackers. In other words, they are theoretical evaluations of the tracker that may be only used for the corresponding tracker.
  - Algorithm-free:
    These types of measures can be applied to every tracker. Various categories can be again defined based on the availability of truths and tracks. When truths and tracks are both available, a large class of performance metrics may be defined after finding an association between the estimated tracks and available truths. In real scenarios, truths are not available and, therefore, the first type is more suitable for simulation environments. In this case, statistical tests are done on the estimated tracking results to check the consistency of the estimates.

It is not possible to judge the quality of a tracker based on a specific group of measures. As shown in [7], it can be observed that the tracker may have provided inaccurate results while some measures show correct and satisfactory performances. Therefore, in a real scenario, all or a mixture of represented metrics from different categories are used in order to evaluate the performance of a tracker and compare its results to the other methods.

IV. Algorithm-Free Metrics

While performance metrics may be developed for individual tracking algorithms and applications, the main focus of this paper is on the algorithm-free metrics. A detailed version of other categories of performance metrics can be found in [9].

Different categories of algorithm-free metrics may be defined based on the availability of truths and tracking results as

- Available truths and tracks:
  There are three classes of metrics summarized as follows:
  1) **Track Cardinality Measures:**
     This metric measures numerical characteristics of obtained results such as the number of confirmed

tracks associated with truths, and number of missed and false tracks.

2) **Time (Durational) Accuracy Measures:** Time performance of estimated tracks is evaluated by this class of metrics. These performance measures provide more information about the persistency of a track such as the track probability of detection.

3) **Accuracy Measures:** This is the most common measure evaluating the closeness of estimated values to the truths. Several measures can be defined based on the type of distance between the set of truths and tracks.

- Available tracks and unknown truths: This case is also very common in real scenarios when there is no information about truths. In this situation, the consistency of tracking results may be checked. The innovation of tracking is used as the main source of information and, then, common statistical tests may be made on the received information. There are also other scoring metrics defined regarding to the innovation that can be found in [5].

When truths and tracks are both available, the first step is to find an association between two given sets. The most convenient way is to use an assignment based algorithm for the data association [14]. Although other distances such as Wsserstein distances [17] have been also used in the literature for track to truth association, the upcoming performance metrics are defined joint with an assignment matrix between truths and tracks. Different techniques for finding the association matrix can be found in [1]. Now, define an $L(k) \times N(k)$ matrix $\bar{C}(k)$ as the association matrix at the k-th time step as

$$\bar{C}_{ln}(k) = \begin{cases} 1 & \text{If the } l-th \text{ truth assigned to the } n-th \text{ track} \\ 0 & \text{Otherwise} \end{cases} \tag{7}$$

where the above matrix is called C if the following conditions are imposed on the rows and columns of $\bar{C}$

$$\begin{aligned} 0 \leq \sum_{l=1}^{L(k)} \bar{C}_{ln}(k) \leq 1 \\ 0 \leq \sum_{n=1}^{N(k)} \bar{C}_{ln}(k) \leq 1 \end{aligned} \tag{8}$$

Given the association matrix, the following performance metrics can be defined for an MTT problem.

A. Cardinality Metrics

Assume the set of all association matrices for the time interval $k \in [k_1, k_2]$ as $\{C(k_1), C(k_1+1), ..., C(k_2)\}$. In addition, each scenario is run for M Monte Carlo runs. A list of cardinality measures can be defined as

- **Number of Valid Tracks (NVT):** A track is validated if it is assigned to only one target and, also, the assigned target is not associated with any other track [15]. Mathematically, the above expression states that the n-th track is validated if the following equalities hold

$$\sum_{l=1}^{L(k)} C_{ln}(k) = 1 \tag{9}$$

The number of valid tracks can be found for different Monte Carlo runs. The final reported value is an average over all values found for every Monte Carlo run.

- **Number of Missed Targets (NMT):** A target is missed if it is not associated with any track. Defining $N^m_{missed}(k)$ as the number of missed targets for the m-th Monte Carlo run and at the k-th step, it is incremented if $\sum_{l=1}^{L(k)} C_{nl}(k) = 0$ [15]. This equality is checked for all confirmed tracks.
- **Number of False Tracks (NFT):** A track is detected as a false one if it is not assigned to any target. $N^m_{false}(k)$, the number of false tracks for the m-th Monte Carlo run and at the k-th step, is incremented if $\sum_{l=1}^{L(k)} C_{nl} = 0$. The average number of false tracks is obtained as [15]

$$N^*_{false}(k) = \frac{1}{M} \sum_{m=1}^{M} N^m_{false}(k) \tag{10}$$

- **Number of Spurious Tracks (NST):** A track is spurious if it is assigned to more than one target. In this case, matrix C can not be used and $\bar{C}$ is utilized instead. Excluding valid tracks from the set of tracks that can be labelled as spurious, spurious tracks are chosen from the set of false tracks that are not associated with any target. If $N^m_{spur}(k)$ denotes the number of spurious tracks for the m-th Monte Carlo run and at the k-th time step, it is incremented once $\sum_{l=1}^{L(k)} \bar{C}_{nl} > 1$ and the n-th track has not been represented as a valid track using matrix C. Here, the inequality is again checked for all confirmed tracks.
- **Measure of Completeness (MOC):** The following equation is given for the measure of completeness [15]:

$$Comp(k) = \frac{N_{val}(k)}{L(k)} \tag{11}$$

where $N_{val}(k)$ denotes the number of valid tracks and, for simplicity, the index of Monte Carlo run has been omitted. The closer the value to one, the better performance of the tracker is achieved. [2].

- **Average Number of Swaps in Tracks (ANST):** In MTT problems, it is probable that different confirmed tracks are assigned to a certain truth. This may happen for crossing targets when targets approach each other. The number of swaps in the tracks is counted for every available truth. Assume NS^m_l as the number of swaps at the m-th Monte Carlo run defined for the l-th target.

[2]It is assumed that targets are point sources and do not occupy several resolution cells.

Note that, the above value can be found by checking the track segment for each truth over all time steps when the target is alive. The track segment can be also obtained by checking the association matrix at each step. The average number of swaps in tracks can be now calculated by [15]

$$NS^* = \frac{1}{ML}\sum_{m=1}^{M}\sum_{l=1}^{L} NS_l^m \tag{12}$$

where L is the total number of truths in the period of the simulation. It is obvious that closer the above value is to 0, more consistently track to target assignment is achieved.

- **Average Number of Broken Tracks (ANBT):**
It is also probable that there is no track assigned to the truth for several time steps. In [15], it is stated that the number of broken tracks is counted at each time step if there is no assigned track to the truth. To do this, the number of broken tracks NB_l^m for the l-th truth and at the m-th Monte Carlo run is found by checking the track segment associated with the truth. Assume k_l as the last time that a track is assigned to the truth. If k_l is smaller than the last appearance time of the l-th target, NB_l^m is assumed to be 1, which states that there is a break in the estimated trajectory of the l-th target. The final value for the average number of broken tracks, NB^* is found by taking an average over all truths and Monte Carlo runs.
- **Track Continuity (TC):**
The continuity measure for the l-th truth is defined as [6]

$$Tc_l^m = \frac{1}{N_l^t}\sum_{n=1}^{N_l^t}\frac{\Delta k_l^n}{\Delta k_l} \tag{13}$$

where N_l^t denotes the total number of tracks assigned to the l-th truth, Δk_l^n is the duration of the n-th track, and Δk_l corresponds to the appearance time of the l-th truth. Note that, when there is no 0 value in the set of assigned tracks to the truth, the continuity measure is simply written as $\frac{1}{N_l^t}$. It can be also observed that the track continuity has the information of breaks and swaps together. Therefore, values closer to 100% represent less broken or swapped tracks.
- **Tracks Redundancy (TR):**
The total number of tracks that are assigned to, at least, one target is found as $N_{valid}^m(k) + N_{spur}^m(k)$. Track redundancy is defined as the ratio of validated tracks and total assigned tracks as [15]

$$TRR^m(k) = \frac{N_{valid}^m(k)}{N_{valid}^m(k) + N_{spur}^m(k)} \tag{14}$$

- **Spurious Track Mean Ratio (STMR):**
The same definition made for the track redundancy can be provided for the ratio of the number of spurious tracks to the total number of tracks as [15]

$$STR^m(k) = \frac{N_{spur}^m(k)}{N_{spur}^m(k) + N_{spur}^m(k)} \tag{15}$$

B. Time Metrics

Some of important time metrics for the performance evaluation of MTT problems are represented as follows:

- **Rate of False Alarm (RFA):**
Rate of false alarm is defined as the number of false tracks per time step [15]. Define $FLR(k)$ as the rate of false alarm at the k-th step. The rate of false alarm can be written in the following way based on the definition for the average number of false tracks [15]

$$FLR(k) = \frac{N_{false}^*(k)}{\Delta t} \tag{16}$$

where $\Delta t = t_{k_2} - t_{k_1}$ is the simulation time.
- **Track Probability of Detection (TPD):**
In the time interval $[k_1, k_2]$, assume k_{first}^l and k_{last}^l as the first and last time that the l-th target is present, respectively. Now, define T_l^* as the duration that the l-th target is assigned to a valid track. This quantity can be calculated by the following equation:

$$T_l^* = \sum_{k'=1}^{k^*} t_{k'} \tag{17}$$

where k' denotes those time steps in which the l-th target is associated with a valid track. The set of the above time steps can be found by referring to the association matrix C. The probability of detection for each target is now found as [10]

$$P_d^l = \frac{T_l^*}{t_{k_{last}^l} - t_{k_{first}^l}} \tag{18}$$

The final track probability of detection is reported by taking the average over all individual probabilities as

$$P_d = \frac{1}{L}\sum_{l=1}^{L} p_d^l \tag{19}$$

- **Rate of Track Fragmentation (RTF):**
It is probable that the tracking algorithm fragments tracks which means the tracker cannot provide a continuous track for the truth. Having the track segment assigned to the l-th truth, the number of changes in the assigned track IDs is defined as TFR_l^m. A smaller value of TFR_l represents a more persistent tracking algorithm. The rate of track fragmentation, TFR^*, is then defined as an average over all Monte Carlo runs and truths [10].
- **Track Latency (TL):**
In MTT problems, the time lag of detecting new or dead targets may be a criterion for comparison among different tracking algorithms. Track latency can be separately defined for above-mentioned events in terms of *confirmed track latency* (CTL), *tentative track latency* (TTL), and *dead track latency* (DTL) [15]. The smaller value the above-represented quantities have, the better target detection capability is achieved by the algorithm. Assume $\lceil_{m,l}^c$, $\lceil_{m,l}^t$, and $\lceil_{m,l}^d$ as the track confirmation latency, tentative track latency, and dead track latency

for the m-th Monte Carlo run and the l-th target, respectively. Above values can be calculated by finding the associations between truths and targets with regard to the association matrix. The final value is either defined as the worst track latencies found for every above-mentioned events [15] or as an average over all Monte Carlo runs and targets [6].

- **Total Execution Time (TET):**
  Computational cost is another important factor that should be taken in performance evaluation of tracking algorithms. While some algorithms such as MHT provide accurate and optimal results, they suffer from the huge computational cost. Therefore, for every tracking algorithm, the total time needed in order to run the tracker is represented as the total execution time. This metric may be really important in practice where trackers are supposed to be applied to the real time problems.

C. Track Accuracy

Assume that the association matrix has been found at the k-th time step and the l-th target is associated with the n-th track. For the m-th Monte Carlo run, the error vector is defined by

$$\mathbf{e}_m(k) = e\left(\mathcal{T}_n(k), \mathcal{L}_l(k)\right) \tag{20}$$

where e is a function to find the error of estimation based on the given truth. Accuracy metrics may be defined in different ways. Metrics may be classified based on how optimistic or pessimistic they are [13]. The other viewpoint is to classify metrics to absolute and relative measures according to the availability of a reference for metric calculation [13]. Each class of metrics has its own advantages and disadvantages, and specific applications. In the following, a summary of the most common accuracy measures are represented from different aforementioned classes. Interested readers can refer to [9] for more details.

- **Root Mean Squared Error (RMSE):**
  RMSE of estimation for the l-th target and n-th associated track can be now found as

$$\mathcal{R}^l(k) = \sqrt{\frac{1}{N_M^l(k)} \sum_{m=1}^{N_M^l(k)} ||\mathbf{e}_m^l(k)||^2} \tag{21}$$

  where $N_M^l(k)$ denotes the number of Monte Carlo runs that the l-th target is detected at the k-th time step. Given values of RMSE at each time step, the following average RMSE is defined for the whole scenario

$$\bar{\mathcal{R}}^l = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{R}^l(k)\right)^2} \tag{22}$$

- **Average Euclidean Error (AEE):**
  Similarly, Average Euclidean Error (AEE) can be defined for the tracking system as

$$\mathcal{E}^l(k) = \frac{1}{N_M^l(k)} \sum_{m=1}^{N_M^l(k)} ||\mathbf{e}_m^l(k)|| \tag{23}$$

  Again, the average AEE can be calculated for the whole scenario using the following equation

$$\bar{\mathcal{E}}^l(k) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{E}^l(k) \tag{24}$$

- **Average Harmonic Error (AHE):**
  It is stated that both RMSE and AEE focus on the large error terms and, indeed, both are pessimistic metrics [13]. Average Harmonic Error (AHE) is instead an optimistic metric concentrating on good error terms. The following equation can be defined for AHE [13]

$$\mathcal{H}^l(k) = \left(\frac{1}{N_M^l(k)} \sum_{m=1}^{N_M^l(k)} ||\mathbf{e}_m^l(k)||^{-1} \right)^{-1} \tag{25}$$

  The average AHE can be also found by taking an average over individual AHEs as

$$\bar{\mathcal{H}}^l(k) = \left(\frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{H}^l(k)\right)^{-1} \right)^{-1} \tag{26}$$

- **Average Geometric Error (AGE):**
  None of the above-defined metrics are balanced. That is, they are either optimistic or pessimistic. Average Geometric Error (AGE) was proposed in [13] as a balanced metric having both characteristics of optimistic and pessimistic metrics together. The following sets of equations are given for average AGE and AGE at each time step

$$\log\left(\mathcal{G}^l(k)\right) = \frac{1}{N_M^l(k)} \sum_{m=1}^{N_M^l(k)} \log\left(||\mathbf{e}_m^l(k)||\right) \tag{27}$$

$$\log\left(\mathcal{G}^l\right) = \frac{1}{K} \sum_{k=1}^{K} \log\left(\log\left(\mathcal{G}^l(k)\right)\right) \tag{28}$$

The previously represented metrics are all absolute measures, which means the metric is not found in comparison to a reference. Relative measures can be more informative due to a more meaningful interpretation of the reported metric that is being computed relative to a given reference. The type of every metric is then determined according to the selection of the reference. The improvement over the predicted states and also the relative improvement to the measurement error can be considered as two references for the definition of relative measures. New metrics can be defined based on the ratio of measures for updated and predicted tracks. Due to space limitations, details of this class of measures are omitted here. More details of relative measures can be found in [9].

V. Performance Evaluation In the Absence of Truths

It is possible to define some measures based on the individual track's data. Assume the interval $[k_1, k_2]$ as the simulation time and $N_{\mathcal{T}}$ as the total number of confirmed tracks in the same period. The following metrics can be defined for each track:

- **Track Life Time:**
  The metric is defined to find the amount of the time that a track has been alive. Note that tracks may be removed after some steps in which they are not associated with any measurement. Tracks with very short life time are more probable to be false.
- **Track Consistency:**
  There are many situations when a track is alive but not associated with any measurement in some time steps. Define track consistency as the number of time steps that a track is assigned to a measurement. Given TC_n^m as the number of time steps in which the track is not associated with any measurement, the defined metric is incremented at each time step if the n-th track at the m-th Monte Carlo run is not assigned to any measurement. Then, the average of the above metric is found as $TC_n^* = \frac{1}{M}\sum_{m=1}^{M} TC_n^m$.

Another class of performance metrics for the unknown truth case can be obtained by considering the measurement innovation. Given the innovation and its covariance for each track as $\mathbf{s}_n^m(k)$ and $P_n^{s,m}(k)$, respectively, statistical tests may be made over the innovation in order to check the consistency of tracking results. χ^2 test is the most well-known statistical test to evaluate the consistency of estimations. The main advantage of χ^2 test is that it is suitable for cases in which parameters of the goal distribution function are estimated from the samples itself [19]. It can be shown that the normalized distance $d^m(k) = \mathbf{s}^m(k)\left(P^{s,m}(k)\right)^{-1}\left(\mathbf{s}^m(k)\right)'$ is χ^2 distributed with n_v as the degree of freedom where n_v is the dimension of innovation [1] [19] [**?**]. The overall distance $d^*(k) = \sum_{m=1}^{M} d^m(k)$ is also χ^2 distributed with $n_v M$ degrees of freedom [**?**]. To evaluate the consistency of the estimations, it should be checked whether the results fall within the defined *confidence region*. The 95% confidence region, $[c_1, c_2]$, is defined as

$$\{c_1, c_2\} : prob\left(c_1 \leq \chi^2_{n_v M} \leq c_2\right) = .95 \tag{29}$$

Values of c_1 and c_2 can be easily found from the χ^2 distribution tables. c_1 and c_2 provide two thresholds in order to evaluate the fitness of estimates at each time step. It is also possible to define different thresholds by changing the confidence probability (.95). The whiteness of the innovation can be also tested using the following statistics [1]

$$\rho(k,j) = \sum_{m=1}^{M} \mathbf{s}^m(k)\left(\mathbf{s}^m(k)\right)' \left[\sum_{m=1}^{M} \mathbf{s}^m(k)\left(\mathbf{s}^m(k)\right)' \sum_{m_2=1}^{M} \mathbf{s}^m(j)\left(\mathbf{s}^m(j)\right)'\right]^{-\frac{1}{2}} \tag{30}$$

The distribution of $\rho(k,j)$ can be found for relatively large values of M. It is stated in [1] that for M large enough, $\rho(k,j)$ is Gaussian distributed. If innovations are also zero mean and white, the mean of $\rho(k,j)$ will be zero and its variance is equal to $\frac{1}{M}$. Therefore, the above test can be used to check the whiteness of innovations.

Besides the χ^2 test, other statistical tests such as Kolomogorov-Smirnov method can be applied to the innovation for performance evaluation. A more detailed discussion of statistical tests for performance evaluation of MTT problems can be found in [9].

VI. Simulation Results

Performance evaluation module is a part of multi-target multi-sensor test-bed developed in McMaster university. Figure 1 shows the main window for the module in the test-bed. There are many options that give enough flexibility to the user. It is possible to choose specific trackers and targets from the list of available options. Also, the time interval of displaying the results may be determined by the user. Three classes of algorithm-free metrics can be also observed in Figure 1. There is also another box for those metrics that may not fall in any of three proposed classes.

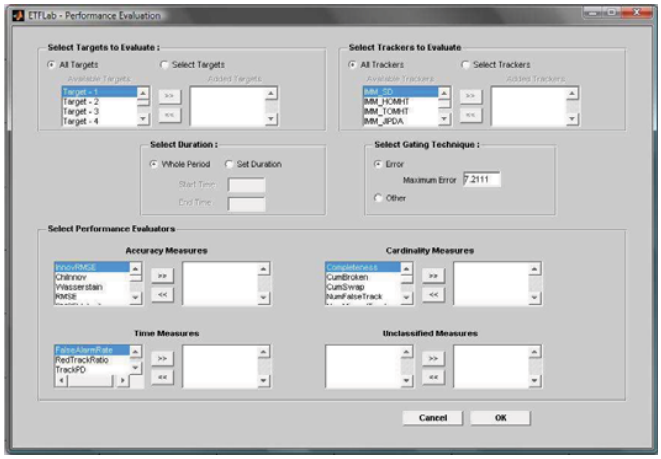

Figure 1. Performance evaluation module in test-bed.

For simulation purposes, a challenging 10-target scenario is chosen. The generated trajectory of targets is shown in Figure 2. It can be observed that targets are all well-separated initially but they approach each other in the subsequent scans. The results of five tracking algorithms are provided for performance evaluation. The following tracking methods have been used for simulations where results are available for 50 Monte Carlo runs:

1) IMM-SD assignment (T1)
2) IMM Hypothesis Oriented MHT (IMM-HOMHT) (T2)
3) IMM Track Oriented MHT (IMM-TOMHT) (T3)
4) IMMJIPDA (T4)
5) IMM Gaussian Mixture PHD (IMMGMPHD) (T5)

Different categories of performance metrics are now found based on the given tracking results. Tables I,II, and II present performance metrics for every tracker separately. There is also a possibility to plot a graph for performance metrics. As an example, the results for the number of valid tracks have been

Table I
ACCURACY MEASURES FOR THE 10-TARGET SCENARIO

Metric	T1	T2	T3	T4	T5
Innovation RMSE (m)	1.34	1.63	1.66	1.56	NaN
χ^2 test	5.11	3.7	3.65	NaN	NaN
Position RMSE (m)	1.15	1.37	1.29	1.18	1.187
Velocity RMSE ($\frac{m}{s}$)	0.0904	0.0903	0.0906	0.0891	0.0907
Position AEE (m)	0.85	0.977	0.932	0.879	0.874
Velocity AEE ($\frac{m}{s}$)	0.0892	0.089	0.0895	0.0879	0.0897
Position AHE (m)	0.242	0.273	0.262	0.26	0.247
Velocity AHE ($\frac{m}{s}$)	0.0861	0.0855	0.0864	0.085	0.0869
Position AGE (m)	0.54	0.603	0.579	0.563	0.553
Velocity AGE ($\frac{m}{s}$)	0.0878	0.0874	0.0881	0.0866	0.0885

shown in Figure 2. Graphs may be also depicted for those metrics that can be computed in different time steps.

Summary tables can be now used in order to compare the proficiency of different trackers. Although people may have different interpretations from the given results, some common conclusions can be drawn from the tables. For example, it can be observed that the track probability of detection is almost one for all trackers. Therefore, all trackers have been able to detect targets efficiently. $T1$ shows more consistent and continuous results compared to other trackers due to the higher continuity measure. This measure can be really informative for this scenario because targets cross each other and tracks may be swapped or broken in the crossing interval. In terms of accuracy, all trackers show relatively the same results. Nevertheless, the cardinality measures are different for trackers where T1 and T2 show the best and worst results, respectively. In other words, it can be concluded that, for this scenario, cardinality and time measures are more informative that the accuracy measures. Therefore, it can be observed again that all performance metrics have to be taken together in order to judge the efficiency of tracking algorithms.

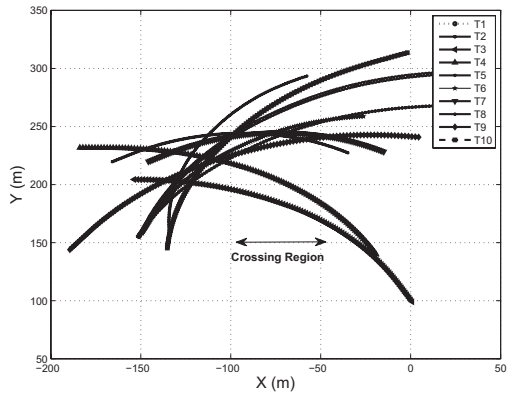

Figure 2. Simulated trajectories for the 10 target scenario.

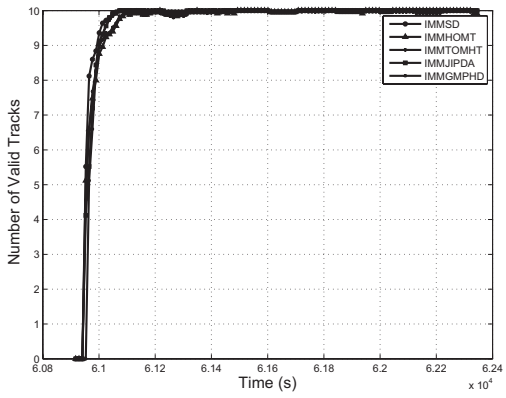

Figure 3. Estimated number of valid tracks for different trackers.

Table II
TIME MEASURES FOR THE 10-TARGET SCENARIO

Metric	T1	T2	T3	T4	T5
RFA	0.00042	0.0085	0.014	0.0007	0.001
RTF	0.975	0.968	0.961	0.975	0.966
TPD	0.973	0.965	0.966	0.968	0.961
RTF	0.656	3.748	1.772	1.996	1.204
Maximum CTL (s)	72	72	72	72	84
Average CTL (s)	36	42	42	44.4	55.2
Maximum DTL (s)	0	0	0	0	0
Average DTL (s)	0	0	0	0	0
Maximum TTL (s)	12	48	48	60	72
Average TTL (s)	1.2	15.6	15.6	27.6	39.6
TET (s)	53.63	47.33	125.35	51.94	32.62

VII. CONCLUSION

An organized classification of performance metrics for MTT problems was presented in this paper. Performance measures were divided into track and sensor-related measures where algorithm-free and algorithm-based metrics were introduced as two sub-categories of track-related measures. Three categories of algorithm-free metrics were represented and several mea-

Table III
CARDINALITY MEASURES FOR THE 10-TARGET SCENARIO

Metric	T1	T2	T3	T4	T5
MOC	0.973	0.965	0.966	0.968	0.961
Average NBT	0.4	5.68	3.48	0.28	0.72
Average NST	1.72	3.04	2.32	5.04	2.8
Average NFT	0.005	0.102	0.172	0.008	0.013
Average NMT	0.267	0.343	0.339	0.314	0.382
Average NST	0.0003	0.0757	0.155	0.003	0.01
Average NVT	9.65	9.57	9.58	9.6	9.53
TC (%)	90.5	67.59	77.16	78.21	84.37

sures of performance were described for each class. Finally, performance measures were applied to a challenging simulated scenario with several targets crossing each other in certain time steps. Results of five well-known trackers were collected and performance metrics were derived for each tracker. It was shown that a combination of metrics from different categories can provide a criterion in order to compare the performance and capability of different tracking algorithms.

REFERENCES

[1] Y. Bar-Shalom, X.R. Li, and T. Kirubarajan, Estimation, "Tracking and Navigation: Theory, Algorithms and Software", John Wiley & Sons, New York, June 2001.

[2] S. Blackman, and R. Popoli, "Design and analysis of modern tracking systems", Artech House, August 1999.

[3] R. M. H. Burton, "A survey of MOP/MOE for maritime command, control, and information systems", *Technical Report*, Department of National Defence, Canada, September 1995.

[4] K. C. Chang, S. Mori, and C. Y. Chong, "Evaluating a multiple-hypothesis multitarget tracking algorithm", IEEE Transactions on Aerospace and Electronic Systems, Vol. 30, No. 2, pp. 578–590, April 1994.

[5] K. C. Chang, and X. Zhao, "A greedy assignment algorithm and its performance evaluation", Proceedings of American Control Conference, Seattle, USA, June 1995.

[6] C. I. Coman, and T. Kreitmair, "Evaluation of the tracking process in ground surveillance applications", Proceedings of the Sixth European Radar Conference, Rome, Italy, September 2009.

[7] S. Coraluppi, D. Grimmett, and P. de Theije, "Benchmark evaluation of multistatic trackers", Proceedings of the 9th International Conference on Information Fusion, Florance, Italy, July 2006.

[8] N. S. Dietrich, "Performance metrics for correlation and tracking algorithms", MSc Thesis, Naval Postgraduate School, Montrerey, California, June 2001.

[9] A. A. Gorji, R. Tharmarasa, and T. Kirubarajan, "Performance Measures for Multiple Target Tracking Problems", Technical Report, Estimation, Tracking and & Fusion Lab, McMaster University, Canada, February 2011.

[10] D. Grimmet, S. Coraluppi, B. R. La Cour, C. G. Hempel, T. Lang, P. A. M. de Theije, and P. Willet, "MSTWG multistatic tracker evaluation using simulated scenario data sets", *Proceedings of the* 11*th International Conference on Information Fusion*, Cologne, Germany, September 2008.

[11] T. Kirubarajan, H. Wang, Y. Bar-Shalom, and K. R. Pattipati, "Efficient multisensor fusion using multidimensional data association", IEEE Transactions on Aerospace and Electronic Systems, Vol. 37, No. 2, pp. 386–400, April 2001.

[12] X. R. Li, and Y. Bar-shalom, "Performance prediction of interacting multiple model algorithm", IEEE Transactions on Aerospace and Electronic Systems, Vol. 29, No. 3, pp. , July 1993.

[13] X. R. Li, and Z. Zhao, "Evaluation of estimation algorithms part I: incomprehensive measures of performance", IEEE Transactions on Aerospace and Electronic Systems, Vol. 42, No. 5, pp. 1340–1358, October 2006.

[14] R. L. Popp, T. Kirubarajan, and K.R. Pattipati, "Survey of Assignment Techniques for Multitarget Tracking", Chapter 2 in Multitarget/Multisensor Tracking: Applications and Advances III, (Y. Bar-Shalom and W.D. Blair, eds.), Artech House, 2000.

[15] R. Rothrock, and O. E. Drummond, "Performance metrics for multiple-sensor, multiple-target tracking", Proceedings of SPIE Conference on Data and Signal Processing of Small Targets, Orlando, USA, July 2000.

[16] D. J. Salmond, and H. Birch, "A particle filter for track-before-detect", Proceedings of American Control Conference, Arlington, VA, USA, June 2001.

[17] 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, August 2008.

[18] I. Simeonova, and T. Semerdjiev, "Specific features of IMM tracking filter design", International Journal of Inoformation and Security, Vol. 9, pp. 154–165, 2002.

[19] M. A. Stephens, "EDF statistics for goodness of fit and some comparisons", Journal of the American Statistical Association, Vol. 69, No. 347, pp. 730–737, September 1974.

[20] J. Vermaak, S. J. Godsill, and P. Perez, "Monte carlo filtering for multi-target tracking and data association", IEEE Transactions on Aerospace and Electronic Systems, Vol. 41, No. 1, pp. , January 2005.

[21] H. Wang, T. Kirubarajan, and Y. Bar-shalom, "Precision large scale air trafic surveillance using IMM/Assignment estimators", IEEE Transactions on Aerospace and Electronic Systems, Vol. 35, No. 1, pp. 255–266, January 1999.

[22] X. Wang, and D. Musicki, "Evaluation of IPDA type filters with a low elevation sea-surface target tracking", Proceedings of the Sixth International Conference on Information Fusion, Queensland, Australia, May 2003.