Noname manuscript No.
(will be inserted by the editor)

Predicting Multiple Target Tracking Performance for Applications on Video Sequences

Juan E. Tapiero · Henry Medeiros · Robert H. Bishop

Received: date / Accepted: date

Abstract This paper presents a framework to predict the performance of multiple target tracking (MTT) techniques. The framework is based on the mathematical descriptors of point processes, the probability generating functional (p.g.fl). It is shown that conceptually the p.g.fls of MTT techniques can be interpreted as a transform that can be marginalized to an expression that encodes all the information regarding the likelihood model as well as the underlying assumptions present in a given tracking technique. In order to use this approach for tracker performance prediction in video sequences, a framework that combines video quality assessment concepts and the marginalized transform is introduced. The multiple hypothesis tracker (MHT) and Markov Chain Monte Carlo (MCMC) data association methods are used as a test cases. We introduce their transforms and perform a numerical comparison to predict their performance under identical conditions.

Keywords Tracking · Multiple Targets · Prediction · Performance

Henry Medeiros
Department of Electrical and Computer Engineering
Marquette University
Tel.: +1-414-288-6186
Fax: +1-414-288-5579
E-mail: henry.medeiros@marquette.edu

Juan E. Tapiero
Department of Electrical and Computer Engineering
Marquette University
E-mail: juan.tapierobernal@marquette.edu

Robert H. Bishop
Department of Electrical Engineering
University of South Florida
E-mail: robertbishop@usf.edu

1 Introduction

The problem of Multiple Target Tracking (MTT) generalizes the notion of single target tracking and state estimation [29,15] to more complex scenarios in which the state of multiple elements must be estimated simultaneously. Since there is no unified theory for single target tracking and estimation, a number of distinct MTT techniques have been proposed, each of which imposes a certain set of assumptions on the problem to make it tractable. Although a number of different approaches have been employed, most of them rely to some extent on Bayesian probabilities. Today the most widely used techniques can be divided into four sets of filters: Extensions of the Bayesian framework for single target tracking [21] (classical examples are Multiple Hypothesis Tracking [22], Joint Probabilistic Data Association filter [8], and Markov Chain Monte Carlo methods [19]), Random Finite Sets framework [14] (examples are Probability Hypothesis Density filters [32], Cardinalized Probability Hypothesis Density filters [31], and Multi-Bernoulli filters [33,10]), Point Processes framework [28] (e.g. Intensity Filter) and heuristic implementations (e.g. Nearest neighbor standard filter) [38]. Many additional examples for each of the frameworks can be found in the literature and, given the renewed interest in the data association problem, particularly in the machine vision community [2,39], that number is expected to continue to increase [12,16].

Despite the great progress made in recent years in the area of MTT and the availability of a large number of algorithms to solve the problem, the choice of MTT technique to be used on any given application is usually made ad-hoc based on the familiarity of the designer with a certain technique. Currently, there are no mechanisms to help determine the chances of suc-

cess of any tracking technique other than applying different methods to a certain problem and comparing them using standard target tracking metrics. But could there be some fundamental characteristics of the problem than can point us towards the selection of one of the techniques as being preferable over all the others? Can the performance of the selected technique be predicted? This paper attempts to answer these questions.

The theory of Point Processes has been used for statistical analysis and estimation of data in many applications [7,18,14,24]. Its connection with classical MTT techniques was first explored in the early developments of Random Finite Sets statistics [4,14] and further developed later by Streit *et al.* [27,26,25]. Point Processes now correspond to a general framework that encompasses the main target tracking approaches previously mentioned. One important set of tools from the point processes framework that relates all the techniques is the probability generating functional (p.g.fl.). In [27], Streit *et al.* uses p.g.fls. to outline in a clear and intuitive manner the way in which different assumptions about the state spaces and measurement spaces for a large set of widely known techniques give rise to different tracking approaches.

This work proposes a framework for predicting the performance of MTT techniques applied to video sequences. The focus on video-based tracking stems from the important applications that multiple target tracking on video sequence has in fields such as robotics, surveillance, video game industry and sports broadcasting [13]. It is also motivated by the possibility of objectively measuring challenging characteristics of the video per se [3], which can give us an implicit understanding of the state and measurement spaces under consideration. This information, in conjunction with the fundamental relationships described by probability generating functionals can be used as a framework for tracking performance prediction. In this work we also comment on the probability generating functional representation for the Markov Chain Monte Carlo data association approach, which had not been introduced in [27], although briefly mentioned in [26].

The remainder of this article is organized as follows. Section 2 introduces the general problem of multiple target tracking from the classical Bayesian point of view using standard set definitions. It then describes the Multiple Hypothesis tracker and the Markov Chain Monte Carlo data association techniques, which we consider pilot test techniques for our contribution. Section 3 presents the theoretical definition of point processes and probability generating functionals, outlining their relationship to the general target tracking problem. Section 4 presents the main conceptual interpretation of the probability generating functionals for multiple target tracking techniques as a transform that encodes all the information provided by a measurement at a given time instant. Section 5 shows the framework that uses visual quality assessment techniques in combination with the introduced transform to obtain a quantity that we cal tracker quality assessment (TQA), which allows us to predict the performance of the tracking mechanism. Finally sections 6 and 7 introduce experimental results as well as conclusions and future work.

2 Multiple Target Tracking Problem

Let us start by introducing a general representation of multi-target state spaces and measurement spaces. Let $\mathcal{S}$ be a general state space for each of the targets $\mathbf{x}(t_k)$. At any given instant of time, the total number of targets $\bar{N}(t_k)$ is unknown. We can designate a region, $\mathcal{R}$, which defines the boundaries of the tracking problem. Given this region we can then add an additional state ϕ to the target state space $\mathcal{S}$ that denotes whether a target is not inside the defined boundary $\mathcal{R}$, then $\mathcal{S}^+ = \mathcal{S} \cup \{\phi\}$ and this is true for each target, giving us a joint state space $\mathcal{S} = \mathcal{S}^+ \times \cdots \times \mathcal{S}^+$ where the product is taken $\bar{N}(t_k)$ times. With this in mind we can represent the set of targets at any given time $t_k > 0$ taking into account that new targets can be born

$$\mathbf{X}(t_k) = \{\mathbf{x}_1(t_k), \mathbf{x}_2(t_k), ..., \mathbf{x}_{n'}(t_k)\} \cup \{\mathbf{b}_1, ..., \mathbf{b}_\nu\}, \quad (1)$$

where $\mathbf{x}(t_k)_{1...n'}$ are the targets that persist from the last instant of time and $\mathbf{b}_{1...\nu}$ are the new born targets.

For the measurement model we have the classical representation $\mathbf{y}_{j,k} = \mathbf{h}(\mathbf{x}_j(t_k), t_k, \mathbf{v}(t_k))$ for the j^{th} target, where $\mathbf{v}(t_k)$ is the measurement noise. We can extend this to a set of $\bar{M}$ measurements that in general can be produced from the targets in the set $\mathbf{X}$ or by false alarms (clutter in the environment) or wrong measurements. This set can be represented as

$$\mathbf{Y}_k = \{\mathbf{y}_{1,k}, ..., \mathbf{y}_{\bar{M},k}\}. \quad (2)$$

The objective of multi-target Bayesian estimation in this case is to estimate the contents of the set $\mathbf{X}(t_k)$ recursively, based on the set of observations $\mathbf{Y}_k$, using the joint transition density for the state $p(\mathbf{X}(t_k)|\mathbf{X}(t_{k-1}))$ and the joint likelihood function $p(\mathbf{Y}_k|\mathbf{X}(t_k))$. We have also the assumptions that are key to Bayesian estimation and inference described for MTT. First, the Markov assumption states that the values in any set of states $\mathbf{X}(t_k)$ are only influenced by the values of the set of states that directly preceded it in time. This implies that the future is independent of the past given

knowledge about the present. In a continuous-discrete setting, we have

$$p(\mathbf{X}(t_{0:k})) = \prod_{i=1}^{k} p(\mathbf{X}(t_i)|\mathbf{X}(t_{i-1}))p(\mathbf{X}(t_0)). \quad (3)$$

We also have the conditional independence of the set of observations, which states that the observation set, $\mathbf{Y}_k$, given the state, $\mathbf{X}(t_k)$, is conditionally independent from the observation and state history, or

$$p(\mathbf{Y}_{1:k}|\mathbf{X}(t_{0:k})) = \prod_{i=1}^{k} p(\mathbf{Y}_i|\mathbf{X}(t_i)). \quad (4)$$

Finally, the estimation process for multiple targets ideally follows the same procedure as the single target case but with the use of the joint densities of all the targets. That is, given the state at time step t_{k-1}, Bayes' theorem is used to determine the joint posterior density at time t_k. This can be achieved in two steps as described below.

Given the motion model and the Bayesian joint posterior density $p(\mathbf{X}(t_{k-1})|\mathbf{Y}_{1:k-1})$ at time t_{k-1}, a time-updated joint density is obtained using the Chapman-Kolmogorov equation:

$$p(\mathbf{X}(t_k)|\mathbf{Y}_{1:k-1}) = \int p(\mathbf{X}(t_k)|\mathbf{X}(t_{k-1})) \times p(\mathbf{X}(t_{k-1})|\mathbf{Y}_{1:k-1})\mathrm{d}\mathbf{X}(t_{k-1}). \quad (5)$$

The observation set $\mathbf{Y}_k$ is then used to update (weigh) the density produced by the time-update step to determine the final joint posterior density at time t_k:

$$p(\mathbf{X}(t_k)|\mathbf{Y}_k) = \frac{p(\mathbf{Y}_k|\mathbf{X}(t_k))p(\mathbf{X}(t_k)|\mathbf{Y}_{1:k-1})}{p(\mathbf{Y}_{1:k})}. \quad (6)$$

The joint posterior density function $p(\mathbf{X}(t_k)|\mathbf{Y}_k)$ encapsulates everything about the set of target states, based on the current set of observations and a priori information. The calculations needed to obtain the exact posterior of this unified estimation are even more challenging than in the single object case. Hence, the different algorithms that have been designed to try to solve them rely on additional assumptions that facilitate implementations of sequential solutions. In the next sections we introduce two approaches that attempt to solve this problem based on different assumptions: multi hypothesis tracking and single scan Markov Chain Monte Carlo data association.

2.1 Multiple Hypothesis Tracking

The Multi Hypothesis Tracker (MHT) is a method for calculating the probabilities of various data association hypotheses. It maintains several hypotheses for each target at each instant of time. In order to do this, this technique enumerates all possible associations over time. As each measurement is obtained, it is classified according to its probability of origin: coming from a previously known target, from a false measurement, or from a new target. The estimation of each possible hypothesis is done through the Kalman filter (for Gaussian transition densities as introduced in the original literature [22]). As additional information (or measurements) is collected, the probabilities of joint hypotheses are calculated sequentially using all the prior knowledge about the system such as the density of unknown targets, probability of detection, and density of false targets. This general technique is usually regarded as a hypotheses-oriented or measurement-to-target ($M \rightarrow T$) data association [30].

2.2 Single Scan Markov Chain Monte Carlo Data Association

Markov Chain Monte Carlo (MCMC) data association is an extension of the Joint Probabilistic data association (JPDA) approach, which was designed to allow a varying number of targets. The JPDA tackles uncertain data association conditions by allowing a target to be updated by a weighted sum of all the measurements within a certain distance threshold of the target. The weights represent the probability that the measurement originates from that particular target. As such, a measurement can contribute to more than one track, and its contribution is weighted according to its association probability. MCMC data association expands on this by considering the space of all possible associations, where each association event may correspond to three possible conditions: deletion, addition (survival) or persistence (move) [19]. The weights are calculated in a manner similar to the JPDA but Monte Carlo methods, such as Metropolis-Hastings sampling [11], are used to integrate over the set and evaluate the probability of each of the three conditions above.

3 Finite Point Processes and Probability Generating Functionals

Finite point processes are usually introduced in the framework of the theory of random measures. Let χ be a topological space (complete, separable, metric). A typical choice for χ is $\mathbb{R}^d$, $d > 0$. The space of sets of points or event space in χ is defined by

$$\varepsilon_\chi = \emptyset \cup \bigcup_{n \geq 1} \chi(n), \quad (7)$$

where $\chi(n)$ is the space of sets of size $n \in \mathbb{N}$, that is $\chi(n) = \big\{\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} | \mathbf{x}_i \in \chi, i = 1, \ldots, n\big\}$. All its elements are assumed to be locally finite and each bounded subset of χ can contain only a finite number of points [24].

Although a point process is regarded as a random (multi)set $\{\mathbf{x}_i\}_i \subset \chi$, it is technically convenient to formally define it as a random measure $\Phi = \sum_i \delta_{\mathbf{x}_i}$ or the mapping

$$\Phi : (\Omega, \mathcal{F}, \mathbb{P}) \to (\varepsilon_\chi, B(\varepsilon_\chi)), \tag{8}$$

where $(\Omega, \mathcal{F}, \mathbb{P})$ is an arbitrary probability space and $B(\varepsilon_\chi)$ denotes the Borel σ-algebra of ε_χ. Hence, if Φ denotes the point process $\{\phi_i\}$, we write $\Phi(A)$ for the number of points ϕ_i that belong to a subset $A \subseteq \chi$; similarly, for suitable functions f on χ, $\int f \, d\Phi = \sum_i f(\phi_i)$.

If Φ is a point process on χ, there exists a unique Borel measure ν on χ such that $\mathbb{E}\Phi(A) = \nu(A)$ for every Borel set A, and more generally $\mathbb{E} \int h \, d\Phi = \int h \, d\nu$ for every positive measurable function h. This measure ν is called the *intensity* of Φ and it completely statistically describes it.

Another descriptor of the point process comes from the finite-dimensional distributions ('fidi') of a random measure ξ that are the joint distributions, for all finite families of bounded Borel sets $A_1, \ldots, A_k$ of the random variables $\xi(A_1), \ldots, \xi(A_k)$, that is the image of the probability measure $\mathbb{P}$, represented as P_Φ [18].

Even though the notation for the spaces presented in Section 2 and the definition of a point process above are seemingly different, the nature of the state space Eq. (2) and measurement sets Eq. (1) can be interpreted as a point process. The change in cardinality (birth/death of targets) at a given time represents a different counting measure $\chi(n)$ in Eq. (7). Furthermore, both share the $\emptyset$ set that accounts for false targets or measurements. Similar arguments can be presented for their probability spaces and their interactions.

3.1 Probability Generating Functional

The information about a point process can be encoded in an algebraic expression called a probability generating functional (p.g.fl). Before introducing the mathematical definition of probability generating functionals, it is important to introduce the most commonly known concept of probability generating functions that is also useful in the context of multiple target tracking.

Given a random variable X on the space (X, B_χ) the probability generating function (p.g.f.) of X is the function defined, for each $z \in \mathbb{R}$, as $\mathbb{E}[z^X]$

$$G(z) = E[z^X] = \sum_{x=0}^{\infty} p(x) z^x. \tag{9}$$

This of course can be extended to multiple dimensions where $(z_1, \ldots, z_n) \in \mathbb{R}^n$. In order to introduce the general mathematical formulation of the probability generating functional (p.g.fl.) we must first introduce $V(\chi)$ the set of B_χ -measurable (test) functions $h : \chi \to \mathbb{R}$ such that $1 - h(x)$ vanishes out of some bounded set and $0 \leq h(x) \leq 1$ for each $\mathbf{x} \in \chi$, with this the p.g.fl. of a general point process Φ on the space χ is defined, for each $h \in V(\chi)$ [24], as

$$\Psi[h] \equiv \Psi_\Phi[h] = \mathbb{E}\left[\exp\left(\int_\chi \log[h(\mathbf{x})] \Phi(d\mathbf{x})\right)\right]. \tag{10}$$

Since the process Φ is defined to be finite on the set where $1 - h(\mathbf{x}) \neq 1$ then it can be written as

$$\Psi[h] \equiv \Psi_\Phi[h] = \mathbb{E}\left[\prod_{i=1} h(\mathbf{x}_i)\right], \tag{11}$$

where $\mathbf{x}_i$ are the points such that $\Phi = \sum_i \delta_{\mathbf{x}_i}$, possibly having repetitions in the (multi)set $\{x_i\}$. In order to realize this expected value, we resort to the fidi of the point process, which allows us to rewrite the p.g.fl as [37]

$$\Psi[h] \equiv \sum_{n \geq 0} \int_{\chi(n)} \prod_{i=1}^{n} h(\mathbf{x}_i) P_\Phi(d\{\mathbf{x}_1, ..., \mathbf{x}_n\}) \tag{12}$$

$$= \sum_{n \geq 0} \frac{1}{n!} \int_{\chi(n)} \prod_{i=1}^{n} h(\mathbf{x}_i) p_n(\mathbf{x}_1, ..., \mathbf{x}_n) d\mathbf{x}_1 ... d\mathbf{x}_n, \tag{13}$$

where the final representation is obtained thanks to the combinatorics interpenetration of a Janossy measure [37] applied to the fidi. This representation can then be extended to joint point processes, where a new process Υ is introduced, with similar characteristics to Φ but on space $\mathcal{Y} \in \mathbb{R}^{d_y}$ (in this application it can be considered as the measurement space and it has a mapping to the state space). Thus, Eq. (13) can be extended to the joint space and defined on $\varepsilon_\chi \times \varepsilon_\Upsilon$ as the product of the random measures

$$\Psi_{\Phi\Upsilon}[g, h] \equiv \sum_{m \geq 0} \sum_{n \geq 0} \frac{1}{m! n!} \int_{\Upsilon^m} \int_{\chi^n} \prod_{i=1}^{m} g(\mathbf{y}_i) \prod_{i=1}^{n} h(\mathbf{x}_i)$$

$$p_{\Phi\Upsilon}(\mathbf{y}_1, ..., \mathbf{y}_m, \mathbf{x}_1, ..., \mathbf{x}_n) d\mathbf{y}_1 ... d\mathbf{y}_m d\mathbf{x}_1 ... d\mathbf{x}_n, \tag{14}$$

where g has the same definition as h as a vanishing test function. Marginalizing this p.g.fl with respect to one process results in the p.g.fl of the other process

$$\Psi_{\Phi\Upsilon}[1, h] = \Psi_\Phi[h] \qquad \text{and} \qquad \Psi_{\Phi\Upsilon}[g, 1] = \Psi_\Upsilon[g]. \tag{15}$$

4 Multiple Target Tracking Transform

The idea of a Multiple Target Tracking transform is inspired by the fact that a generating function is an algebraic tool for encoding combinatorial data [20]. With this in mind, we can claim that the given p.g.fs and p.g.fls encode all the combinational information of the finite point process they represent. If it is assumed that a tracking technique is a joint finite point process [27], then its p.g.fl representation encodes all the information pertinent to target and measurement existence and the set of assumptions encoded within the technique. It is also a fact that the test functions in which a p.g.fl is evaluated are complex numbers (like any classical transform), which in this case represent a space of "existence" for each target (on a z-transform for example it represents a discrete time delay). This means that if a target exists there is a complex number representing its existence and the same occurs for measurements.

In order to use p.g.fls for the estimation of the state of the targets, it is necessary to evaluate all these complex variables (that vary on dimension too) as can be seen in [26], which is the equivalent of finding the inverse transform for classical techniques. In our case, we want to predict the performance of the technique, so obviously we want to avoid performing all the steps in the estimation processes. Instead, we rely on the p.g.fls, which, if properly interpreted, provide us some insights on how we can achieve a new simplified representation for multi-target tracking performance prediction.

4.1 Brief Review of Important Probability Generating Functionals

In order to make this article self-contained, we introduce some important p.g.fl definitions. The p.g.fl of a Poisson point process representing the clutter in the measurements is given by

$$\Psi_C^{PPP}[g] = exp\left(-\Lambda + \Lambda \int_{\mathbf{Y}} g(\mathbf{y}) p_\Lambda(\mathbf{y}) d\mathbf{y}\right), \tag{16}$$

where Λ is the mean number of clutter points in $\mathcal{Y}$ (measurement space) and $p_\Lambda(\mathbf{y})$ is the normalized intensity function. For tracking applications these functions are usually assumed constant in the window of observation.

The p.g.fl of a Bayes-Markov filter with probability of miss-detections is given by (see section IV.A in [27])

$$\Psi_{BMD}[g,h] \equiv \int_{\mathbf{S}} h(\mathbf{x})\mu(\mathbf{x}) \times \left(a(\mathbf{x}) + b(\mathbf{x}) \int_{\mathbf{Y}} g(\mathbf{y}) p(\mathbf{y}|\mathbf{x}) d\mathbf{y}\right) d\mathbf{x}. \tag{17}$$

The different coefficients represent the probability of a target state $\mu(\mathbf{x})$, likelihood of the measurement $p(\mathbf{y}|\mathbf{x})$, probability of detection $b(\mathbf{x})$ and probability of miss-detection $a(\mathbf{x})$.

When target births are taken into consideration, the data-induced targets are represented by a process with p.g.fl $\Psi_{BMD}^{Data}[g,h]$, which has the same functional form as $\Psi_{BMD}[g,h]$, but with the target density $\mu(\mathbf{x})$ replaced by a prior distribution $\xi(\mathbf{x})$. When evaluated under linear Gaussian assumptions, this process corresponds to a regular Kalman filter.

4.2 Marginalized Transforms

The proposed simplified representation is based on the application of Eq. (15) to the p.g.fl of the tracking techniques under consideration. In this work we use the MHT representation, which was previously introduced by Streit *et al.* in [27]

$$\begin{aligned}\Psi[g,h_1,...,h_{n+m}] &= \Psi_C^{PPP}[g] \\ &\times \prod_{i=1}^{n} \left[1 - \chi_i + \chi_i \Psi_{BMD(i)}[g,h_i]\right] \\ &\times \prod_{j=1}^{m} \left[1 - \gamma_j + \gamma_j \Psi_{BMD(j)}^{Data}[g,h_{n+j}]\right],\end{aligned} \tag{18}$$

where χ_i is the probability of existence, and γ_j is the probability of birth from acquired data respectively.

As part of our contribution, we introduce the p.g.fl for the MCMC approach, which has the same form as the p.g.fl for MHT. It differs from the former, however, in the way the coefficients χ and γ are calculated, as explained in Sections 2.1 and 2.2. The marginalized transform for both the MHT and MCMC techniques is therefore given by

$$\begin{aligned}\Psi[g] = \Psi_C^{PPP}[g] &\left(1 - \chi_i + \chi_i \Psi_{BMD}[g]\right)^n \\ &\times \left(1 - \gamma_i + \gamma_i \Psi_{BMD}^{Data}[g]\right)^m,\end{aligned} \tag{19}$$

and it represents all the information encoded on each measurement. In other words, it provides us a measure of the amount of information that can be obtained from a set of measurements on a given instant of time.

5 Framework for Tracker Performance Prediction in Videos

The marginalized transform gives us general information on the expected performance of a given tracking techniques according to the measurements obtained. In order to apply this information specifically to videos, we introduce a framework that uses information encoded

within the video itself since this will affect the measurements in a unique manner for each different scenario. Our objective is to use the concept of video quality assessment with the tracking technique in the role of quality observer.

5.1 Visual Quality Assessment

Research in objective image quality assessment seeks to design quantitative measures with the capability to automatically predict perceived image quality [35]. Once developed, an objective image quality metric can be applied to an extensive range of practical uses. These applications include image acquisition, compression, communication, displaying, printing, restoration, enhancement, analysis, and watermarking.

In this work we apply two different and widely used techniques which present different features. First we have the blind/referenceless image spatial quality evaluator also known as BRISQUE [17], which is a natural scene statistic-based distortion-generic blind/no-reference image quality assessment model that operates in the spatial domain. We also incorporate the multi-scale structural similarity (MS-SSIM) [36], which is an extension of the structural similarity (SSIM) index [34]. SSIM uses structural distortion as an estimation of the perceived visual distortion. In order to determine structural distortion, SSIM utilizes means, variances, and the covariance of a reference and a given image. The outputs of SSIM and MS-SSIM have similar numerical values over time but MS-SSIM has smaller absolute variation since it takes in account more details of the image. It is important to note that these techniques were chosen because they perform quality assessment on the spatial dimension of the image, where the tracking application is performed.

5.2 Tracker Quality Assessment

The process of incorporating the MTT transform into a quality assessment framework consists of five steps (shown in Figure 1). First, each frame is weighted by the quality score obtained using BRISQUE. At the same time, the detector that is going to be used for the tracking application is applied to each frame, obtaining real measurements. These measurements are then used to numerically compute the MTT transform in Eq. (19) for each technique under consideration. In order to do this, each measurement is assumed to represent a target with all the possibilities provided by the technique[1]. The BRISQUE-weighted image is then modified by applying the weighting obtained by the transform, but only in the pixel regions where measurement were obtained. Finally the MS-SSIM is used to compare the weighted images from the current frame to the previous one. Figure 2 shows one example of the weights generated by the proposed framework for one snapshot of the TUD-Stadtmitte dataset (see Figure 3). Darker regions correspond to higher weight values. As the figure indicates, the proposed approach focuses on regions where both target detections and image changes occur. The tracker quality assessment (TQA) is the cumulative difference of the output provided by MS-SSIM frame to frame.

In practice, for usual applications of MS-SSIM the higher its value the higher the quality of the image under consideration, since the reference is the image with perfect quality. The comparison here occurs on a frame-by-frame basis, but the interpretation remains the same, the larger the total change observed the higher the quality of the tracking technique, which corresponds to more efficient information encoding.

6 Experimental Results

In its simplest form, the evaluation of the tracker quality assessment is straightforward. Given a video sequence, we can simply apply the tracker quality assessment framework and obtain a quantity that predicts the expected performance of the technique. In order to carry out this experiment we need to run the tracking techniques on the video sequence to have a real performance measure that can be compared with the prediction. For this work, we use the optimal subpattern assignment (OSPA) metric [23] since it has been widely used as one of the main performance metrics for non-labeled multiple target tracking applications. The implementation of the MHT tracker used in this paper is based on the work of Antunes *et al.* [1] and the MCMC data association is available in the toolbox by Särkkä *et al.* [9]. Both methods were extended or modified to carry out centroid tracking on videos with a constant velocity model [13]. In all of our experiments, we carried out 15 Monte Carlo simulation trials for each tracking scenario.

The video sequences analyzed are the widely utilized and publicly available VSPETS 2003 INMOVE soccer dataset[2], the TUD-Stadtmitte dataset[3], and the 2009

[1] Although this is a relatively strong assumption, it provides reasonable results. Making this assumption weaker is subject of future work, as explained in more detail in our concluding remarks.

[2] `ftp://ftp.cs.rdg.ac.uk/pub/VS-PETS/`

[3] `https://motchallenge.net/vis/TUD-Stadtmitte`

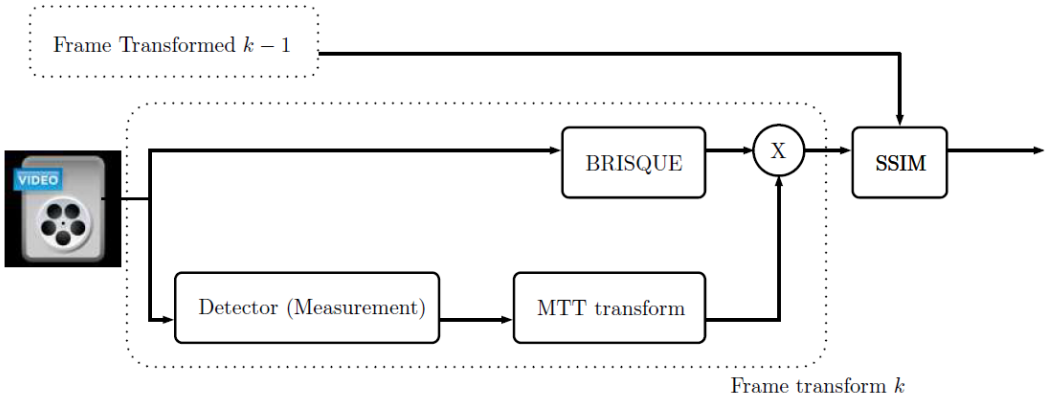


Fig. 1: Framework for the application of the MTT transform combined with visual quality assessment.

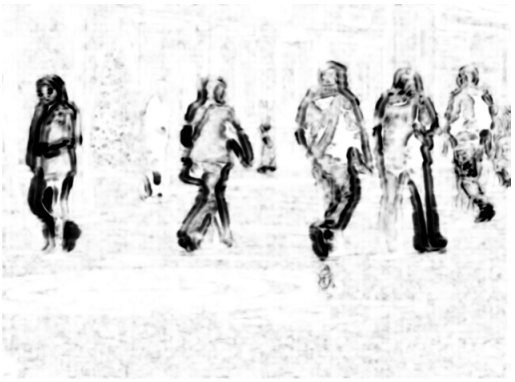


Fig. 2: Example of the tracker quality assessment weights by generated by the proposed framework. The method focuses on areas where detections occur.

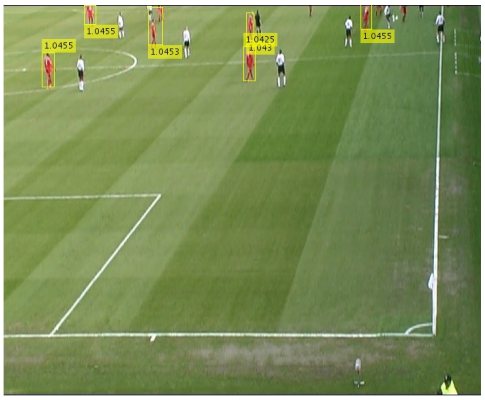
Table 1: Sets of assumption for the MCMC tracker in the soccer scenario. c_d is the clutter or false alarm density, p_d is the probability of death, and p_b is the probability of birth.

MCMC Assumption sets			
1	$c_d = 1/1000$	$p_d = 0.547$	$p_b = 0.1$
2	$c_d = 1/240$	$p_d = 0.8$	$p_b = 0.1$
3	$c_d = 1/1000$	$p_d = 0.9$	$p_b = 0.1$
4	$c_d = 1/100$	$p_d = 0.9$	$p_b = 0.1$
5	$c_d = 1/3$	$p_d = 0.547$	$p_b = 0.8$

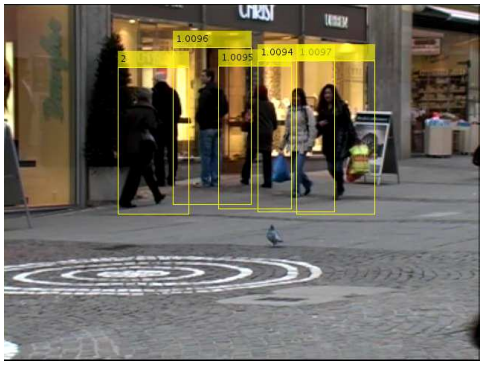
BAHNHOF sequence which corresponds to a moving camera scenario[4]. In the VSPETS dataset, we use a red color detector to obtain the centroids of one of the teams (Liverpool), which in general provides very accurate measurements. For the TUD-Stadtmitte and the BAHNHOF datasets, the targets are pedestrians and the detections were carried out with using HOG [5]. In these scenarios there is significantly more clutter and the detection accuracy is not as high, particularly for the moving camera case. The main assumption for the transform evaluation is that a target is present wherever a measurement is present (at marginalization) and a transform value is calculated for each of them. Then for each frame, if there are q measurements, we obtain q transform values that are superposed since the measurement space is unique. This also implies that the values for m and n in Eq. (19) are assumed to be equal to the number of measurements obtained at each frame.

[4] `https://data.vision.ee.ethz.ch/cvl/aess/dataset/`

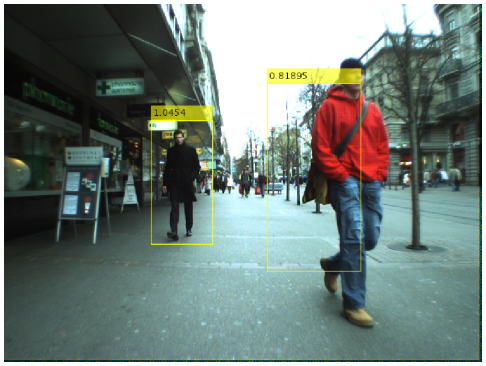
The tracker quality assessment framework was used to evaluate each tracker using different sets of basic assumptions, such as varying the false alarm intensities, the probability of detection and/or probability of birth. Although every parameter of the tracking algorithms could be varied for our evaluation, from the model and measurement covariance to the birth densities or the motion model itself, for simplicity and conciseness we limit our analysis to the parameters shown in Table 1 and in Figure 4. For the soccer scenario, after extensive experimentation, it was found that changing the false alarm density affects this specific MHT implementation the most, and hence that was the parameter chosen for evaluation. In the case of the MCMC technique, varying a single parameter does not significantly affect the tracker performance given its ability to more thoroughly explore the hypothesis space. Hence, several different assumption sets were evaluated, as shown in Table 1.



(a) Soccer scenario VSPETS 2003 INMOVE.



(b) Pedestrian scenario TUD-Stadtmitte.



(c) Moving camera scenario 2009 BAHNHOF.

Fig. 3: Snapshots of the datasets analyzed.

6.1 MTT Transform Coefficient Calculation

Before presenting numerical results, it is necessary to introduce the expression for the calculation of the coefficients γ and χ since this is where the chosen techniques, MHT and MCMC, differ in terms of their p.g.fl and marginalized transform as was mentioned in section 4.2. For a general linear (or approximately linear) Gaussian assumption for likelihood and motion models, the pseudocode to calculate the marginalized transform for the techniques under evaluation is given in Algorithm 1. In summary, the procedure consists of taking a set of measurements and first calculating the Bayes-Markov filter using the standard Kalman filter calculations of the mean, covariance and innovation (lines 5-9) for existing and new targets. Lines 11 and 12 are the calculation of the Bayes-Markov filter from Eq. (17) for existing and new targets, with probabilities of detection and miss-detection a_x and b_x, respectively. After evaluating the probabilities of existence and of birth in lines 13 and 14, line 15 computes Eq. (19) for each target assumed to be present at a measurement (clutter is included in χ and γ). The total transform is computed by adding over the measurement space over the m iterations of the algorithm.

Algorithm 1 Marginalized transform calculation.

1: **while** video is running **do**
2: $\boldsymbol{Y} \leftarrow$ set of measurements for current frame
▷ $\boldsymbol{Y}$ is an $l \times m$ matrix
3: $\Psi = 0$
4: **for** $i = 1$ to m **do**
5: $\boldsymbol{x} = \boldsymbol{Y}_i$
▷ Assume a target is present at each measurement
6: $\boldsymbol{x}_+ = A \cdot \boldsymbol{x}$
7: $Y_+ = H \cdot \boldsymbol{x}_+$
8: $P_+ = A \cdot P_0 \cdot A^T + Q$
9: $S = H \cdot P_+ \cdot H^T + R$
10: $BMD = a_x \cdot \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}_+, P_+) + b_x \cdot \mathcal{N}(Y_i|Y_+, S)$
▷ Account for detection and model effects
11: $S_{data} = H \cdot P_{birth} \cdot H^T + R$
12: $BMD_{data} = a_x \cdot \mathcal{N}(x_+|M_{birth}, P_{birth})$
$+b_x \cdot \mathcal{N}(Y_+|Y_i, S_{data})$
▷ Effects of birth density with normal distribution with mean M_{birth} and covariance P_{birth}
13: Evaluate χ according to Eq. (20) or Eq. (22).
14: Evaluate γ according to Eq. (21) or Eq. (23).
15: $\Psi = \Psi + (1 - \chi + \chi * BMD)^m$
$\cdot (1 - \gamma + \gamma * BMD_{data})^m$
16: **end for**
17: **end while**

6.2 Multiple Hypothesis Tracking

For the MHT we need to introduce the expression for the calculation of the coefficients γ and χ. Using the Gaussian assumption and the expressions for hypothesis evaluation from the MHT we have

$$\chi_{MHT} = \frac{1}{c_d} \cdot \frac{\nu! F!}{\bar{Q}!} \cdot e^{-p_b} \cdot p_d \cdot \left(e^{-c_d}\right)^F, \tag{20}$$

$$\gamma_{MHT} = \frac{1}{c_d} \cdot \frac{\nu! F!}{\bar{Q}!} \cdot e^{-c_d} \cdot \left(e^{-p_b}\right)^\nu, \tag{21}$$

where F is the number of false alarms, ν is the number of new targets, $\bar{Q}$ is the number of targets, and again p_b and p_d are the probabilities of birth and death, respectively and c_d is the false alarm density [22,6]. All the

former quantities are calculated using random sampling according to the appropriate distribution: binomial distribution for ν, and Poisson distribution for F. $\bar{Q}$ is approximated by the number of measurements for each frame. It is important to remember that in this case the probability of detection and the innovation probability density function are already taken into account inside the Bayes-Markov filter portion of the marginalized transform expression.

6.3 Markov Chain Monte Carlo

For the MCMC approach, the evaluation has a different nature, since it depends on three conditions mentioned in Section 2.2. In this case we have

$$\chi_{MCMC} = (1 - p_b) \cdot p_d \cdot \tau \cdot C, \tag{22}$$

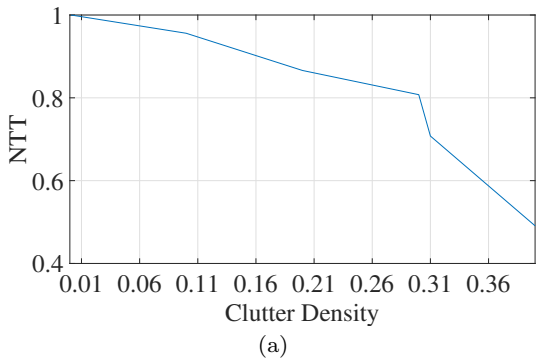
$$\gamma_{MCMC} = p_b \cdot (1 - p_d) \cdot C, \tag{23}$$

where p_b is the probability of birth, p_d is the probability of death, τ is the target prior, and C is the clutter prior for a target. The value for each of these variables is calculated by performing a small Metropolis-Hastings sampling [11] using the different assumptions present in the implementation. To calculate those values, we used the same criteria presented by Särkkä *et al.* in [9]. The value of C is sampled from a Poisson distribution with density c_d and it is equal to the inverse of the surveillance volume (in this case, the image area) if the target is said to be a false alarm, otherwise it is one. τ depends on sampling from a given target representing a false alarm or existing target, and it is obtained by sampling a Poisson distribution with intensity p_b and assigning the value of one minus the inverse of the surveillance volume.

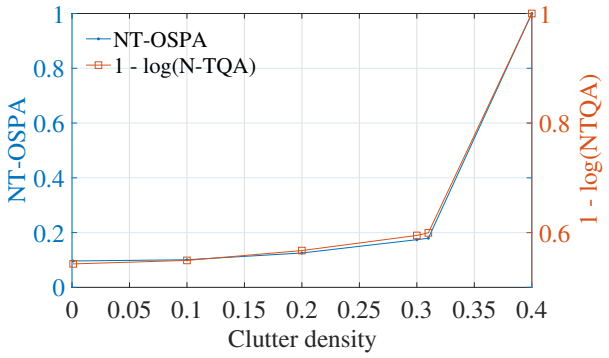
6.4 Numerical Results

In order to evaluate our hypothesis, we analyze three sets of metrics. Total OSPA values are provided to demonstrate the actual performance of the techniques using the ground truth. It is computed by accumulating the OSPA value of each video frame. Total MTT transform represents the aggregated value of the transform evaluated for each frame. Finally, the TQA represents the output of the framework presented in Section 5.2. All quantities have been normalized by their largest value in order to facilitate visualization and to facilitate the comparison with the tracker performance.

In the soccer scenario, it can be observed in Figure 4.b that the normalized TQA (i.e., $1 - \log(\text{NTQA})$) does an excellent job predicting the performance for the MHT technique, considering that smaller normalized



(a)



(b)

Fig. 4: MHT results for the soccer scenario. NTT stands for normalized total MTT, NT-OSPA is the normalized total OSPA, and NTQA is the normalized TQA.

total OSPA values (NT-OSPA) correspond to better overall performances. In this case the normalized total MTT transform (NTT) also performs well (Figure 4.a), but it is important to remember that it only takes in account the detection and not the characteristics of the video itself. For the MCMC tracker, we can observe in Figure 7 a small variation on the performance prediction for the first three sets of assumptions which is also reflected on the actual OSPA. In general for this scenario the TQA framework performs very well, mostly due to the high quality of the measurements and the low number of false alarms and video changes, given the stationary camera.

In the pedestrian dataset, despite the more challenging scenario, which includes significant partial occlusions, as well as the different detector, the TQA framework can still predict the performance of the MHT tracker very accurately as shown in Figure 5. Although the TQA in Figure 5.a does not follow the OSPA as closely as in the soccer scenario, it still reflects its growth very accurately. Although the behavior of the NTT for the pedestrian dataset shown in Figure 5.b might seem identical to that for the soccer dataset (Fig. 4.b), it

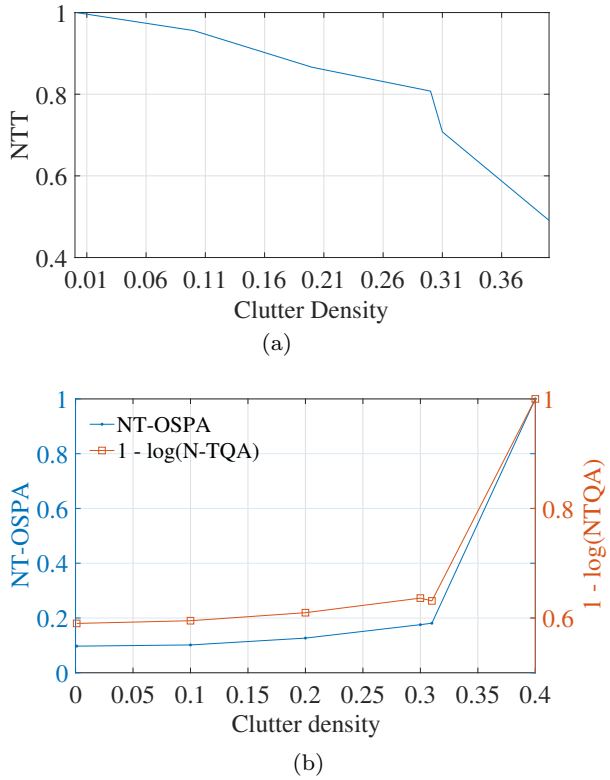


Fig. 5: MHT results for the pedestrian scenario.

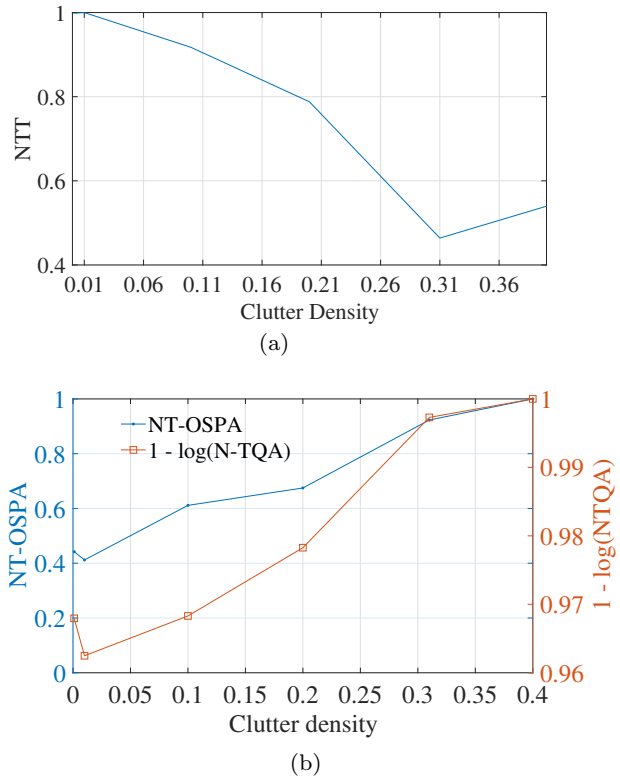


Fig. 6: MHT results for the moving camera scenario.

should again be noted that these are normalized values. The absolute values of both transforms differ by one order of magnitude. The maximum NTT value for the soccer dataset is approximately 15,000 whereas for the pedestrian dataset it is close to 1,000. This difference reflects the significantly more challenging conditions seen in the second scenario. For the MCMC method, it can be observed in Figure 8 that although the TQA still reflects the decreased OSPA, it varies slightly more slowly. In this case, since the targets are quite large with respect to the background, the frame-to-frame changes in the image tend to impact the TQA more than in the soccer scenario. In this case, the NTT (Figure 8.a) follows the decrease in OSPA more closely.

The moving camera scenario presents further challenges for the evaluated tracking techniques and that can be seen on the TQA framework. For the MCMC method, it can be observed that the variation of the TQA value is smaller than the actual OSPA variation and the trends are less precise than in the stationary camera scenario (Figures 6.b and 9.b). On the other hand, the transform presents a more accurate estimate in this case, since it is not as affected by the even larger frame to frame changes (Figures 6.a and 9.a). For the sets of assumptions with relatively low OSPA (i.e., good performance), the TQA provides an accurate estimate of performance. Although the TQA prediction is not as accurate for higher values of OSPA, it still provides a good estimate of the expected performance when used in conjunction with the transform values.

7 Conclusions and Future Work

The mathematical framework of finite point processes allows for the introduction of novel concepts that can be used to produce compact representations of MTT techniques. These representations can be used to obtain more information about the nature of these techniques and devise applications that go beyond simple target tracking. We used these concepts to present a new framework that allows us to predict the performance of MTT techniques without performing tracking.

The MTT transform gives us an insight on how the different assumptions of MTT techniques affect the way in which the information content of the measurements is used. Although the MTT transform by itself gives us information about the effective use of the measurements, it is not a complete prediction since the scenario in which tracking occurs also affects the performance.

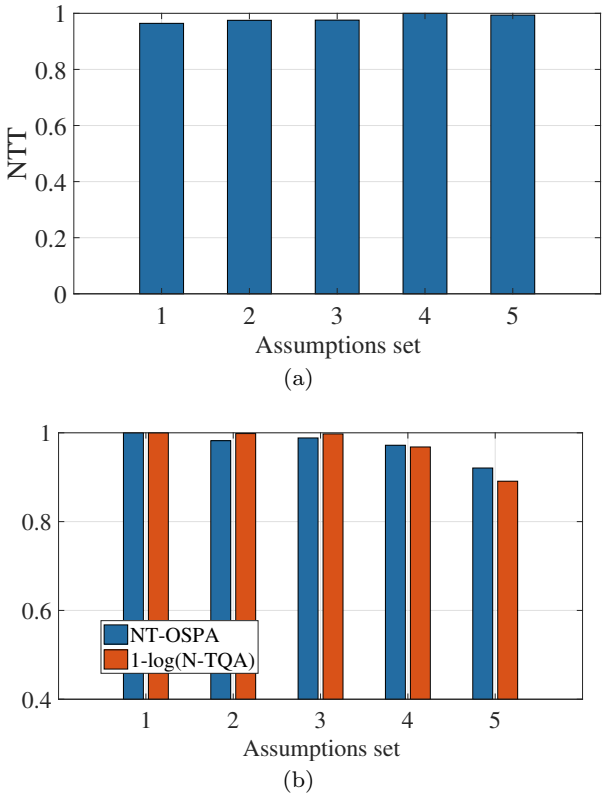


Fig. 7: MCMC results for the soccer scenario.

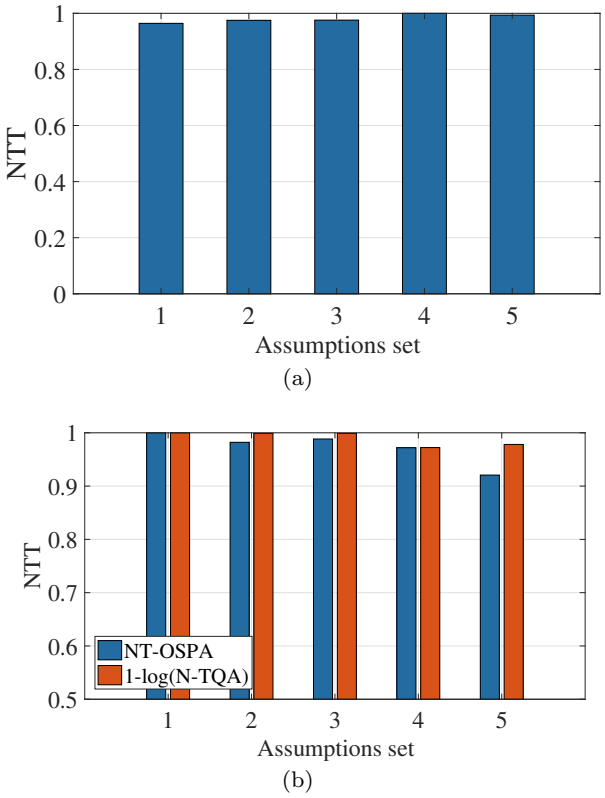


Fig. 8: MCMC results for the pedestrian scenario.

Visual quality assessment techniques can therefore be successfully integrated with the proposed transform to give a more accurate performance prediction that takes into account the problems of quality present in the video sequences and its dynamic nature. Our experiments demonstrated that the proposed framework can successfully predict the tracking performance of two different tracking approaches, MHT and MCMC, as measured by the OSPA metric under different conditions.

In the future, we would like to extend our method to make more accurate use of the measurements available at each image frame. In our current approach, each measurement is associated with a potential target. One possible strategy to mitigate this assumption would be to perform local measurement clustering so that targets are associated with clusters of measurements instead. Another area of potential improvement is the possibility of applying different weights to the output of BRISQUE and of the MTT to account for effects such as highly dynamic backgrounds. This should allow us to further improve the accuracy of the TQA for dynamic background scenarios such as when the camera is moving.

References

1. Antunes, D.M., de Matos, D.M., Gaspar, J.: A library for implementing the multiple hypothesis tracking algorithm. arXiv preprint (2011)
2. Cancela, B., Ortega, M., Penedo, M.G.: Multiple human tracking system for unpredictable trajectories. Machine Vision and Applications **25**(2), 511–527 (2014)
3. Chikkerur, S., Sundaram, V., Reisslein, M., Karam, L.J.: Objective video quality assessment methods: A classification, review, and performance comparison. IEEE Transactions on Broadcasting **57**(2), 165–182 (2011)
4. Cressie, N., Laslett, G.: Random set theory and problems of modeling. SIAM Review **29**(4), 557–574 (1987)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893 vol. 1 (2005). DOI 10.1109/CVPR.2005.177
6. Danchick, R., Newnam, G.: Reformulating Reid's MHT method with generalised murty k-best ranked linear assignment algorithm. In: IEEE Proceedings on Radar, Sonar and Navigation, vol. 153, pp. 13–22. IEE (2006)
7. EmBree, J.D.: Spatial temporal exponential-family point processes for the evolution of social systems. Ph.D. thesis, University of California Los Angeles (2015)
8. Fortmann, T.E., Bar-Shalom, Y., Scheffe, M.: Sonar tracking of multiple targets using joint probabilistic data association. IEEE Journal of Oceanic Engineering **8**(3), 173–184 (1983)

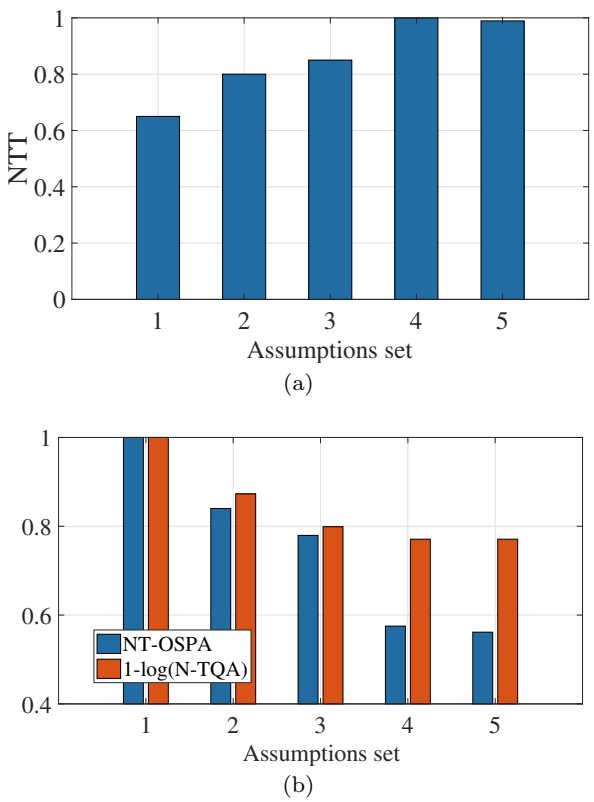


Fig. 9: MCMC results for the moving camera scenario.

9. Hartikainen, J., Särkkä, S.: RBMCDAbox-Matlab toolbox of rao-blackwellized data association particle filters. documentation of RBMCDA Toolbox for Matlab V (2008)
10. Hoak, A., Medeiros, H., Povinelli, R.J.: Image-based multi-target tracking through multi-bernoulli filtering with interactive likelihoods. Sensors **In press** (2017)
11. Kroese, D.: Handbook of Monte Carlo methods. Wiley, Hoboken, N.J (2011)
12. Leal-Taixá, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint (2015)
13. Maggio, E., Cavallaro, A.: Video tracking: theory and practice. John Wiley and Sons (2011)
14. Mahler, R.P.: Statistical multisource-multitarget information fusion. Artech House, Inc. (2007)
15. Medeiros, H., Holguin, G., Shin, P.J., Park, J.: A parallel histogram-based particle filter for object tracking on simd-based smart cameras. Computer Vision and Image Understanding **114**(11), 1264–1272 (2010)
16. Milan, A., Leal-Taixé, L., Reid, I.D., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. CoRR **abs/1603.00831** (2016). URL `http://arxiv.org/abs/1603.00831`
17. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012)
18. Moyal, J.: The general theory of stochastic population processes. Acta mathematica **108**(1), 1–31 (1962)
19. Oh, S., Russell, S., Sastry, S.: Markov chain monte carlo data association for general multiple-target tracking problems. In: 43rd Conference on Decision and Control, vol. 1, pp. 735–742. IEEE (2004)
20. Pemantle, R., Wilson, M.C.: Analytic Combinatorics in Several Variables, vol. 140. Cambridge University Press (2013)
21. Pulford, G.W.: Taxonomy of multiple target tracking methods. IEEE Proceedings on Radar, Sonar and Navigation **152**(5), 291–304 (2005). ID: 1
22. Reid, D.B.: An algorithm for tracking multiple targets. IEEE Transactions on Automatic Control **24**(6), 843–854 (1979)
23. Ristic, B., Vo, B.N., Clark, D., Vo, B.T.: A metric for performance evaluation of multi-target tracking algorithms. IEEE Transactions on Signal Processing **59**(7), 3452–3457 (2011)
24. Spinelli, B.M.: Statistical inference for stable point processes. Ph.D. thesis (2012)
25. Streit, R.: The probability generating functional for finite point processes, and its application to the comparison of phd and intensity filters. Journal of Advances in Information Fusion (2013)
26. Streit, R.: Saddle point method for JPDA and related filters. In: Information Fusion (Fusion), 2015 18th International Conference on, pp. 1680–1687. IEEE (2015)
27. Streit, R., Degen, C., Koch, W.: The pointillist family of multitarget tracking filters. arXiv preprint (2015)
28. Streit, R.L.: Poisson Point Processes: Imaging, Tracking, and Sensing. Springer Science and Business Media (2010)
29. Tapiero, J.E., Bishop, R.H.: Bayesian estimation for tracking of spiraling reentry vehicles. AIAA guidance, navigation, and control (GNC) conference (2013)
30. Tinne, D.: Rigorously bayesian multitarget tracking and localization. Ph.D. thesis (2010)
31. Vo, B., Vo, B., Cantoni, A.: The cardinalized probability hypothesis density filter for linear gaussian multi-target models. In: 40th annual conference on Information sciences and systems, pp. 681–686. IEEE (2006)
32. Vo, B.N., Ma, W.K.: The gaussian mixture probability hypothesis density filter. IEEE Transactions on Signal Processing **54**(11), 4091–4104 (2006)
33. Vo, B.T., Vo, B.N., Cantoni, A.: The cardinality balanced multi-target multi-bernoulli filter and its implementations. IEEE Transactions on Signal Processing **57**(2), 409–423 (2009)
34. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
35. Wang, Z., Sheikh, H.R., Bovik, A.C.: Objective video quality assessment. The handbook of video databases: design and applications pp. 1041–1078 (2003)
36. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. pp. 1398–1402. IEEE (2003)
37. Westcott, M.: The probability generating functional. Journal of the Australian Mathematical Society **14**, 448–466 (1972)
38. Yeddanapudi, M.K.: Estimation and data association algorithms for multisensor-multitarget tracking. Ph.D. thesis (1996)
39. Zuriarrain, I., Mekonnen, A.A., Lerasle, F., Arana, N.: Tracking-by-detection of multiple persons by a resample-move particle filter. Machine Vision and Applications **24**(8), 1751–1765 (2013)