Bayesian multiple person tracking using probability hypothesis density smoothing

S. Hernández\textsuperscript{a}, P. Sallis\textsuperscript{b}

\textsuperscript{a}Laboratorio de Procesamiento de Información Geoespacial. Universidad Católica del Maule. Talca, Chile.
\textsuperscript{b}Geoinformatics Research Centre. Auckland University of Technology. Auckland, New Zealand.

Abstract

We present a PHD filtering approach to estimate the state of an unknown number of persons in a video sequence. Persons are represented by moving blobs, which are tracked across different frames using a first-order moment approximation to the posterior density. The PHD filter is a good alternative to standard multi-target tracking algorithms, since overrides making explicit associations between measurements and persons locations. The recursive method has linear complexity in the number of targets, so it also has the potential benefit of scaling well with a large number of persons being tracked.

The PHD filter achieves interesting results for the multiple person tracking problem, albeit discarding useful information from higher order interactions. Nevertheless, a backward state-space representation using PHD smoothing can be used to refine the filtered estimates. In this paper, we present two smoothing strategies for improving PHD filter estimates in multiple person tracking. Results from using PHD smoothing techniques in a video sequence shows a slight gain in the cardinality estimates (meaning the number of persons in a particular video frame), but good performance in the individual location estimates.

1. Introduction

A Bayesian method for the unknown number of targets with unknown association hypotheses has been formulated using point processes and random finite sets theories, under the name of the Probability Hypothesis Density (PHD) filter \cite{1}. The method solves the troublesome multi-target estimation problem by approximating the complete posterior distribution of the filtering density by the first–order moment of a Poisson process. Alike the Kalman filter recursion, the PHD filter approach uses all observations from the past in order to produce instantaneous estimates of the number of targets and their locations. Moreover,
the PHD recursion can be efficiently computed in closed form using a Gaussian mixture representation or by means of stochastic integration using sequential Monte Carlo methods, so it is suitable for visual tracking applications [2].

Pedestrian counting and tracking is a challenging computer vision task, with applications in surveillance and video monitoring. Analyzing the size of a crowd along with the dynamics of the group and its members, has the potential benefit of providing real-time detection of anomalies or events of particular interest. However, because of the complexity of extracting meaningful information from single or multiple cameras, the scope and availability of multiple target tracking techniques for crowd analysis has been restricted to constrained environments and calibration conditions [3].

Traditional target tracking algorithms for pedestrian tracking relies on intra-frame and inter-frame association hypotheses, which relates image measurements to predicted person locations [4]. In order to compute association hypotheses one has to make assumptions which are usually hard to satisfy in real environments, and specially difficult in crowded scenarios. Furthermore, occlusion reasoning and persons merging and splitting into groups, leaves a full posterior distribution on the number of persons and the association hypotheses being intractable [5].

Although the filtering approach provides a fairly accurate way to calculate an instant estimate of the state of a dynamic system, we might expect an improvement if we incorporate more information in the production of the estimate. Rather than considering only the past and current observations, the accuracy of the filtered estimate can be improved by also taking into account future observations [6]. This procedure is widely known as smoothing, and recent research has been undertaken on producing smoothed estimates for the PHD filter [7, 8].

In this paper, we consider unsupervised top-down Bayesian detection and tracking of multiple persons in crowded environments using the PHD filter approach. Even though the method is well suited for tracking a large number of persons observed in clutter which might come from illumination changes, the PHD approximation only holds for tracking scenarios where the signal-to-noise ratio is sufficiently high that a target can be well represented by the observed features [9]. Unfortunately, most of the state the art image processing techniques for person detection would require supervised learning techniques that are not well suited for real time applications [10], relies on multiple cameras [11] or computationally expensive appearance models [12].

More specifically, we propose a PHD smoothing approach for the problem of person tracking in crowded environments. Firstly, background segmentation is used to a generate foreground mask. Secondly, a simple 2D segmentation technique using ground plane information is used to perform person detection. Thirdly, the PHD filter is used to recursively estimate the number of persons and their locations. Fourthly, PHD smoothing is used to refine the instantaneous estimates. A schematic diagram of the procedure for performing detection and tracking is shown in Figure 1.

The contributions of this paper can be briefly summarized as:
1. PHD filtering and smoothing are described in Section 3 and 4 respectively. Two different state estimation techniques for the PHD filter and smoother are described. The EM algorithm and Bayesian estimation using a Gibbs sampler are summarized, and the Bayesian method is shown to be more appropriate to the person tracking problem.

2. A suitable performance metric for quantitative assessment of the estimation procedure for multiple person tracking is described in Section 5.

3. A method to perform person tracking in crowded environments using a single static camera using PHD filtering is proposed in Section 6. Each person is assumed to move independently of each other, but no restrictions are made about their trajectories and velocities. The method can deal with clutter originated from errors of the person detection technique and illumination changes.

4. We propose to use smoothing as a method to overcome some drawbacks of the PHD filter approach, such as the bias caused by missed detections originated by short occlusions. A first example is given in Section 6.1 which is used to depict the benefit of smoothing over filtering. A second example is given in Section 6.2 which uses ground truth information to evaluate the proposed approach.

Section 2 presents a summary of the application of the PHD filter to visual tracking. An introduction to the PHD filter is given in Section 3 and the sequential Monte Carlo implementation is also presented.

2. Related work

Detection and tracking of a moving person in a video can be achieved by means of comparing the difference between the current image from a reference frame. This technique is widely known as background subtraction, where the reference frame is usually termed the ‘background model’ [13]. The background model is a representation of the scene without moving parts, and the complexity of level of the model depends on the specific scenario. A basic background subtraction technique can use a unique image as the background model, however this technique easily fail when having small changes of luminance or in the
geometry settings. The output of the background subtraction step is a set of connected regions of pixels belonging to the foreground, and are widely known as ‘blobs’. Each region has pixels that forms an ellipse or a bounding box that can be tracked from frame to frame. Features of the connected regions are detections that can then be taken as noisy observations for a tracking system [14].

Tracking multiple humans is a challenging application because of the difficulty of generating a similarity function for a person using pixel information. Quantifying the information of a group of pixels using a person detection system can be potentially intractable, if we consider all possible orientations and occlusions. Early works for person detection considered vertical histograms where the head of the people can be distinguished, but this method is not robust in case of occlusion. More recent works have considered person detection using supervised learning by means of cascades of descriptors [15], requiring careful training and testing.

The application of the PHD filter to tracking multiple trajectories from features points in sequences of optical images was described in [16]. More recently, the sequential Monte Carlo (SMC) implementation of the PHD filter was applied to the problem of tracking multiple groups of persons in video [17] . Observations were taken from the moments of the blobs, and morphological operators were used to reduce the level of clutter in the system. The method was then compared with a Gaussian mixture implementation which explicitly accounts for birth, death and survival of targets [18]. The authors also provided a data–driven method for initializing the spatial density of birth and death in a scene.

The PHD filter was also used for tracking faces, people and vehicles using color based change detection in [19]. Since the PHD filter approach avoid computing associations between targets and estimated tracks, graph matching was proposed as a post-processing step for handling the data association problem. The authors reported improved accuracy of the algorithm in cluttered images.

An extension to tracking 3D object locations from multiple cameras have been proposed in [20]. The method is able to handle occlusions being present at a single camera, by fusing information from multiple cameras using the PHD filter. Further developments in the application of the PHD filter in visual tracking has been done by considering more data–driven approaches for designing birth and death proposals using scene information in [21] and [22].

3. PHD filter

The problem of performing joint detection and tracking of multiple objects has a natural interpretation under the theory of Poisson point processes [23]. In this case, a model-based approach for detection and tracking of multiple objects can be achieved by using the expectation of a random counting measure. Since a Poisson point process is invariant under transformations, such as thinning, superposition and random translations, the posterior distribution can be also approximated by a Poisson point process [24]. This property becomes extremely useful in visual tracking, where targets may randomly appear, disappear merge
or split, leaving the number of targets to be modeled as a non-stationary discrete random variable [25].

Random Finite Sets (RFS) are more general stochastic models for the analysis of spatial patterns and a finite point process are random functions which describes realizations of sets containing a finite number of elements. A model for tracking multiple objects can perform filtering on a set-valued state \( X_k \), given the history of set-valued observations \( Z_{1:k} \). The set-valued approach allows a time-varying number of objects to appear and disappear with no particular order (the density is invariant to permutations of the members of the set), avoiding explicit data association. Furthermore, when using a Poisson spatial model of the new born targets and clutter, it is also possible to determine the expected number targets using the intensity measure of the resulting Poisson process [26, 27].

The instances of the two RFS \( X_k = \{x_1, x_2, \ldots, x_n\} \) and \( Z_k = \{z_1, z_2, \ldots, z_m\} \) represents a set of targets and observations respectively. Bayesian filtering equations are constructed in a similar fashion as their single target filtering counterparts. In this case the RFS filtering and update equations can be written as follows:

\[
p(X_k | Z_{1:k-1}) = \int p(X_k | X_{k-1}) p(X_{k-1} | Z_{k-1}) \delta X_{k-1} \tag{1}
\]

\[
p(X_k | Z_{1:k}) = \frac{p(Z_k | X_k) p(X_k | Z_{1:k-1})}{p(Z_k | Z_{k-1})} \tag{2}
\]

The probability hypothesis density (PHD) \( D(\cdot) \) is defined as the first-order moment or intensity function of a point process with permutation invariant density \( p(\{x_1, \ldots, x_n\} | Z_{1:k}) = j_n(x_1, \ldots, x_n) \). The PHD repackages the family of such densities into a single function that specifies the probability of finding a target \( x \) in a neighborhood of \( \{x_1, \ldots, x_n\} \), such that:

\[
D(x) = \sum_{n=0}^{\infty} \frac{1}{n!} \int j_n(x, x, x_1, \ldots, x_n) \, dx_1, \ldots, dx_n \tag{3}
\]

A recursive formula for the filtering densities is given by:

\[
D_{k|k-1}(x_k) = \int \left[ \pi_s(x_k | x_{k-1}) p(x_k | x_{k-1}) + \gamma_{k|k-1}(x_k | x_{k-1}) \right] D_{k-1|k-1}(x_{k-1}) \, dx_{k-1} + b_{k|k-1}(x_k) \tag{4}
\]

\[
D_{k|k}(x_k) = L_z(x_k) D_{k|k-1}(x_k) \tag{5}
\]

Where:
\[
L_z(x_k) = \left[ 1 - \pi_d(x_k) + \sum_{z \in \mathcal{Z}_k} \frac{\pi_d(x_k) p(z|x_k)}{\lambda_c c_k(z) + D_k(z)} \right]
\]

\[
D_k(z) = \int \pi_d(x_k) p(z|x_k) D_{k|k-1}(x_k) \, dx_k
\]

And:

\[
b_{k|k-1}(x) : \text{Spontaneous birth probability}
\]

\[
\gamma_{k|k-1}(x|x') : \text{Probability of target spawning}
\]

\[
p(x|x') : \text{Single target Markov transition density}
\]

\[
p(z|x') : \text{Single target likelihood function}
\]

\[
\pi_s(x) : \text{Probability of target survival}
\]

\[
\pi_d(x) : \text{Probability of target detection}
\]

\[
\lambda_c : \text{Average number of Poisson false alarms per unit volume}
\]

\[
c_k(z) : \text{Spatial distribution of false alarms}
\]

The number of targets \( N_{k|k} \) is calculated as the integral of the PHD \( D(\cdot) \) or intensity function of the dynamic point process:

\[
N_{k|k} = \int D_{k|k}(x) \, dx \tag{6}
\]

Algorithm 1 describes the particle approximation to the PHD recursion as given in [28].

3.1. State estimation for the particle PHD filter

Being the first-order moment of a multi-target posterior distribution, the PHD represents the intensity function of an information-theoretic best-fit Poisson process approximation. The integral of that function over the field gives the estimated number of targets, and the positions of targets can be estimated from the peaks of the same function.

In the SMC implementation of the PHD filter, Monte Carlo samples are used to represent the intensity function, so a larger number of particles are used in areas where targets are more likely to exist. Assuming that we have sample from the posterior PHD distribution, clustering methods can be used for estimating the targets states. K-means and the Expectation-Maximization (EM) algorithms are the main approaches for state estimation for the PHD filter [29]. The total number of targets corresponds to the total particle mass, so target states are computed by clustering particles and using the centroids of each cluster. Furthermore, the authors in [29] also incorporated track continuity in the particle PHD filter by using validation techniques in the state estimation.
Algorithm 1 Particle PHD filter

**Require:** \( k \geq 1 \land \{w^i_{k-1}, x^i_{k-1}\}_{i=1}^{L_{k-1}} \)

1: Step 1: Prediction step
2: for \( i = 1, \ldots, L_{k-1} \) do
3: Sample \( \tilde{x}^i_k \sim q_k(\cdot|\tilde{x}^i_{k-1}, Z_k) \)
4: Compute predicted weights \( \tilde{w}^i_{k|k-1} = \frac{\phi(\tilde{x}^i_k, Z_k)}{q_k(\tilde{x}^i_k|\tilde{x}^i_{k-1}, Z_k)} w^i_{k-1} \)
5: end for
6: for \( i = L_{k-1} + 1, \ldots, J_k \) do
7: Sample \( \tilde{x}^i_k \sim p_k(\cdot|Z_k) \)
8: Compute predicted weights for the new born particles
9: \( \tilde{w}^i_{k|k-1} = \frac{1}{J_k} \frac{\gamma(\tilde{x}^i_k)}{p_k(\tilde{x}^i_k|Z_k)} \)
10: end for
11: Step 2: Update step
12: for all \( z \in Z_k \) do
13: Compute \( C_k(z) = \sum_{j=1}^{L_{k-1}+J_k} \psi_{z,k}(\tilde{x}^j_k) w^j_{k|k-1} \)
14: end for
15: for \( i = 1, \ldots, L_{k-1} + J_k \), update weights do
16: \( \tilde{w}^i_k = \left[ \nu(x) + \sum_{z \in Z_k} \frac{\psi_{z,k}(\tilde{x}^i_k)}{\tilde{C}_k(z)} \right] \tilde{w}^i_{k|k-1} \)
17: end for
18: Step 3: Resampling step
19: Compute the total mass \( \tilde{N}_{k|k-1} = \sum_{j=1}^{L_{k-1}+J_k} \tilde{w}^j_k \)
20: Resample \( \{\tilde{w}_k^i, \tilde{x}_k^i\}_{i=1}^{L_k} \)
21: return \( \{w^i_k, x^i_k\}_{i=1}^{L_k} \)

7
Figure 2: Positions and velocities of 3 target states are represented by the particle PHD filter. Monte Carlo samples are tightly clustered around target states, making it difficult to compute the model parameters using the EM algorithm.

K-means is an iterative algorithm which separates the data points into K partitions and estimates the centers of each partition [30]. Each data point is associated to only one of the clusters, and the algorithm iterates until some utility function is minimized. On the other hand, the EM algorithm performs probabilistic classification using a finite mixture model. In this case, the probability of each data point belonging to a particular class is evaluated using a normalized allocation vector to all latent classes.

Since the PHD filter assumes low observation noise, parametric estimation using EM can be difficult. All data points would potentially be tightly clustered around their centers, introducing numerical instability in the calculation of the variances [31]. Furthermore, having only access to a re-sampled particle approximation, could also produce a mismatch between model complexity and the amount of available data.

Maximum likelihood approaches for parametric estimation suffer from local minima and over-fitting, as well dependency on the starting point. Bayesian approaches can be used instead in order to overcome the problems of deterministic estimation using limited data [32]. Figure 3 shows the EM algorithm and Bayesian estimation of a bivariate Gaussian mixture model using the Gibbs sampler. Both methods are implemented with the same number of iterations and the estimated parameter is plotted at each step. It is easy to see the dependence of the EM algorithm on the starting point, while the Gibbs sampler is able to explore the parameter space.

In parametric state estimation for the particle PHD filter, we represent the target states as mixture model with parameters \( \Theta_k = (\theta_1^k, \ldots, \theta_N^k) \) and mixing proportions \( \pi_k = (\pi_1^k, \ldots, \pi_N^k) \), and each particle is sampled from the mixture distribution. The mixture density is a weighted linear combination of all the components of the mixture, with known parametric form. Estimation consists of estimating the values of the weights for all components, as well as the specific parameters for each one of the components. The number of components is usually estimated from data, using information theoretic methods such as the Akaike Information Criteria or the Bayesian
Figure 3: Bivariate Gaussian mixture model estimated with the EM algorithm and the Gibbs sampler. After 100 iterations, the Gibbs sampler has explored different configurations of the parameters (Gaussian mean parameter being represented by red crosses), while the EM algorithm estimates are more dependent on the initial setup.

Information Criteria [33]. When performing state estimation for the PHD filter we already have an estimate of the number of targets, which can instead be used as the expected number of clusters.

\[ p(x_i|\Theta_k) = \sum_{j=1}^{N_k} \pi_j p(x_i|\theta_j^k) \]  

The Bayesian approach for finite mixture modeling uses priors on the parameters, usually called hyper-parameters and the likelihood of the available data with respect to the mixture density. Similarly to the EM algorithm, allocation of the data instances to the mixture components is treated as a missing variable, but in this case a density over the unknown allocations is used to perform inference. Estimation can be performed using Markov Chain Monte Carlo methods such as the Gibbs sampler [33], where the available data corresponds to the particle approximation to the PHD. After the update step of the PHD filter at time \( k \), the available data is a sample from the posterior PHD distribution which is taken as the input for the parametric Bayesian state estimator.

4. PHD smoothing

The PHD filter algorithm provides an approximation to the expectation or first-order moment of the intensity measure of a Poisson point process. The method has the property of being able to explicitly model the birth and deaths of targets, as well as clutter and mis-detections, which can also be subject to spawning or merging. This model-based approach can be appealing in multiple tracking systems where the data association step is non-trivial or cannot be optimally solved.

An alternative solution for improving the PHD filter instantaneous estimates is to perform smoothing or retrodiction. Filtered estimates of the individual target states and the posterior cardinality distribution can be considerably improved by considering a higher data frame than the history of observations. More specifically, PHD filtering can be extended to smoothing and is expected to correct the abrupt changes on the
estimated number of targets and their states, that originate from errors propagated by the filtered distributions.

Let $X_k = \{x_1, \ldots, x_{n_k}\}$ be a set target states in and $Z_{1:T}$ a collection of set-valued measurements collected up to time $T \geq k$. The smoothed PHD can be written as follows:

$$D_{k|T}(x) = \int p(\{x\} \cup X_k | Z_{1:T}) \delta X_k$$  \hspace{1cm} (8)

Accordingly, the smoothed number of targets can then be written as:

$$N_{k|T} = \int D_{k|T}(x) dx$$  \hspace{1cm} (9)

As with the standard linear and non-linear smoothing equations, the PHD smoothing problem might be approached by using fixed-interval smoothing, fixed-lag smoothing or fixed-point smoothing. The algorithms presented here are not dependent on the data interval size, so they can be implemented under each one of these schemes. Notice that, since the PHD is only available for non-ordered sets, the full PHD smoothing distribution $p(X_{1:k} | Z_{1:T})$ is not available, so only the marginal PHD smoothing $D_{k|T}(x)$ in Equation 8 can be approximated. Sections 4.1 and 4.2 describes two possible approximations.

### 4.1. Forward-Backward PHD smoother (FB-PHD)

Following a similar approach to the particle forward-backward smoother, Nandakumar et al. developed a Forward-Backward PHD (FB-PHD) smoother [7] based on a physical-space approach [34].

$$p(X_k | Z_{1:T}) = \int p(X_k, X_{k+1} | Z_{1:T}) \delta X_{k+1}$$  \hspace{1cm} (10)

$$= \int p(X_{k+1} | Z_{1:T}) p(X_k | X_{k+1}, Z_{1:T}) \delta X_{k+1}$$  \hspace{1cm} (11)

$$= p(X_k | Z_{1:k}) \int \frac{p(X_{k+1} | Z_{1:T}) p(X_{k+1} | X_k)}{p(X_{k+1} | Z_{1:k})} \delta X_{k+1}$$  \hspace{1cm} (12)

A particle approximation to the smoothing multi-target density can be written as:

$$\int_B D_{k+1|T}(x) dx = E[ \sum_{x_{k+1} \in B} 1_B(x_{k+1}) ]$$  \hspace{1cm} (13)

$$\approx \sum_{i=1}^{L_{k+1}} 1_B(x_{k+1}^i) w_{k+1}^i | T$$  \hspace{1cm} (14)

Algorithm 2 describes a sequential Monte Carlo approximation to the FB-PHD smoother.
Algorithm 2 Forward-Backward PHD smoother

1: Forward pass.
2: for $k = 0, \ldots, T$ do
3:   Perform SMC to get particles and weights \{${x}_k^i, w^i_k$\}$_{1 \leq i \leq L_k}$.
4: end for
5: Choose $w^i_T|T = w^i_T$.
6: Backward pass
7: for $k = T - 1, \ldots, 0$ do
8:   for all $i \in \{1, \ldots, L_k\}$ do
9:     $\mu_{k+1|i}^i = b_{k+1|i}(x_{k+1}^i) + \sum_{l=1}^{L_k} w^i_k \left[ \pi_s(x_k^l)p(x_{k+1}^i|x_k^l) + \gamma_{k+1|i}(x_{k+1}^i|x_k^l) \right]$.
10: $w^i_{k|T} = w^i_k \left[ \sum_{j=1}^{L_k} w^j_{k+1} \pi_s(x_k^j)p(x_{k+1}^i|x_k^j) \right] / \mu_{k+1|i}^i$.
11: end for
12: Compute smoothed estimated number of targets $\hat{N}_k|T = \sum_{i=1}^{L_k} w^i_{k|T}$.
13: Normalize \{${w}^i_{k|T}$\}$_{1 \leq i \leq L_k}$ to get \{${\hat{N}}_k|T$\}$_{1 \leq i \leq L_k}$.
14: end for

4.2. Two-Filter PHD smoother (TF-PHD)

Another approach for PHD smoothing can be achieved by means of the two-filter formula [35]. In this case, the PHD filter has to be combined with the output of a backward information filter, which propagates the posterior distribution of the random counting measure $N_k|T$ from Equation 9 to be represented by the following factorization:

$$p(X_k|Z_{1:T}) = p(X_k|Z_{1:k-1}, Z_k|T)$$  \hspace{1cm} (15)
$$= p(X_k|Z_{1:k-1}) \frac{p(Z_k|T|X_k)}{p(Z_k|T|Z_{1:k-1})}$$  \hspace{1cm} (16)
$$\propto p(X_k|Z_{1:k-1}) p(Z_k|T|X_k)$$  \hspace{1cm} (17)

where the backward information $p(Z_k|T|X_k)$ filter can be written as:

$$p(Z_{k+1:T}|X_k) = \int p(Z_{k+1:T}, X_{k+1}|X_k) \delta X_{k+1}$$  \hspace{1cm} (18)
$$= \int p(X_{k+1}|X_k)p(Z_{k+1:T}|X_{k+1}) \delta X_{k+1}$$  \hspace{1cm} (19)

The SMC approximation for the backward predicted smoother can then be written as Algorithm 3.

5. Performance Evaluation

In the case of multi-target estimation, performance evaluation is not as straightforward as in the single target case, where the consistency is analyzed by means of the residuals of the estimator. In this case, the number of targets has to be estimated
Algorithm 3 Two-Filter PHD smoother

1: Forward pass.
2: for $k = 0, \ldots, T$ do
3:  Perform SMC to get particles and weights $\{x_{k}^{i}, w_{k}^{i}\}_{1 \leq i \leq L_{k}}$.
4: end for
5: Choose $w_{k|T}^{i} = w_{k}^{i}$.
6: Backward pass
7: for $k = T - 1, \ldots, 0$ do
8:  for all $i \in \{1, \ldots, L_{k}\}$ do
9:   $\psi_{l}^{k} = \sum_{h=1}^{J_{k}} \pi_{d}(x_{l}^{h}) p(z|x_{l}^{h})$
10:  $L_{z}(x_{l}^{k}) = 1 - \pi_{d}(x_{l}^{k}) + \sum_{z \in Z_{k}} \pi_{a}(x_{l}^{k}) p(z|x_{l}^{k}) + \psi_{l}^{k}$
11:  $\alpha_{k+1|k}^{i} = \sum_{l=1}^{J_{k}} w_{k|l} p(x_{l+1|k}^{i} | x_{l}^{k})$
12:  $w_{k|T}^{i} = L_{z}(x_{l}^{k}) \left[ \sum_{j=1}^{L_{k+1}} \frac{w_{k|T}^{j} \pi_{a}(x_{l}^{j}) p(x_{l+1|k}^{i} | x_{l}^{j})}{\alpha_{k+1|k}} + 1 - \pi_{a}(x_{l}^{k}) \right]$ 
13: end for
14: Compute smoothed estimated number of targets $\hat{N}_{k|T} = \sum_{i=1}^{L_{k}} w_{k|T}^{i}$.
15: Normalize $\{w_{k|T}^{i}\}_{1 \leq i \leq L_{k}}$ to get $\{\frac{\hat{N}_{k|T}}{L_{k}}\}_{1 \leq i \leq L_{k}}$.
16: end for

alongside with the individual states, and erroneous associations between targets and observations or targets being closely spaced might lead to different estimation results [36]. Evaluating the performance under different configurations between estimated states and ground truth is challenging because of the ambiguity of evaluating all possible associations. The ambiguity of the estimation performance is shown in Figure 4 for a hypothetical multi-target scenario with two different estimators.

Because of the combinatorial nature of the data association problem, the consistency of the estimator cannot be correctly captured by the MMSE criteria. For that reason, different distance metrics have to be used in order to evaluate the performance of multi-target filters and smoothers.

5.1. Optimal Sub-Pattern Assignment (OSPA) metric

In order to override the problems of the standard distance metrics for set-valued observations, Hoffman and Mahler [37] proposed a construction based on the Wasserstein distance. The Wasserstein distance has a long tradition in theoretical statistics for comparing probability distributions on metric spaces. Intuitively speaking, the Wasserstein distance estimates the cost of turning one probability distribution into another distribution with probably different total mass, by means of the minimal average distance between the two sets. The associated cost is calculated as an optimal-assignment problem for all $m \times n$ transportation matrices $C$, that satisfies $C_{i,j} \neq 0 \forall i, j$ and:

\begin{align*}
\end{align*}
Figure 4: Multi-target estimation ambiguity. 3 frames with different number of targets (ground truth plotted in red) and estimated states (plotted in blue). At frame $k$ the first multi-target estimator in Figure 4(a) has detected 3 of 4 targets, and for the same frame the second estimation in Figure 4(b) detected a different multi-target configuration. The same situation occurs at each frame, so it is not clear which one of the multi-target estimators is closer to the ground truth.

\[
\sum_{j=1}^{n} C_{i,j} = \frac{1}{m} \text{ for } 1 \leq i \leq m \\
\sum_{i=1}^{m} C_{i,j} = \frac{1}{n} \text{ for } 1 \leq j \leq n
\]

The Optimal Mass Transfer (OMAT) metric of order $p$ for non-empty sets $X$ and $Y$ is defined as:

\[
d_p(X, Y) = \min_C \left( \sum_{i=1}^{m} \sum_{j=1}^{n} C_{i,j} d(x_i, y_j)^p \right)^{\frac{1}{p}}
\]
The magnitude order parameter $p$ makes the OSPA metric become more sensitive to outliers as $p$ increases, making the metric more focused on location errors. On the other hand, the cut-off parameter $c$ penalizes cardinality errors instead of wrongly placed locations. Because it is a consistent metric for multi-target error estimation, we use the OSPA\(^1\) metric defined in Equation 23 for performance evaluation.

6. PHD filter and smoother for person tracking and counting

Instead of using an explicit person detection system, we use a PHD filter approach to estimate the locations of an unknown number of persons. A constant velocity model is used as a generative model for the movement of a single person. The forward model calculates the new position of a person using a velocity vector that remains nearly constant in magnitude and direction. Let $x_k = [x_k, y_k]^T$ be the transpose of a 2-dimensional position of a person in the image plane and $\dot{x}_k = [\dot{x}_k, \dot{y}_k]$ its velocity.

In a state-space representation the state vector of a person is written as an augmented vector $x_k = [x_k, \dot{x}_k]$. A linear mapping is used to model the dynamic behavior of a person, and Gaussian noise $w_k$. The position of a single person at the discrete time $k$ can be written as:

$$x_k = F x_{k-1} + w_k$$

$$w_k \sim \mathcal{N}(0, \Sigma_{x_k})$$

where $F$ is a linear transformation matrix in which $dt$ represents the sampling time:

$$F = \begin{bmatrix} 1 & 0 & dt & 0 \\ 0 & 1 & 0 & dt \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The observations $y_k = [y_k, y_k]$ only contain information about the position of a person, so velocity has to be estimated from previous measurements [40]. The velocity is related to the object position as $\dot{x}_k = (x_k - x_{k-1})/dt$ for each sampling interval $dt$. However, since the PHD filter does not perform inter-frame person association, velocity is sampled from a zero-mean Gaussian prior distribution $\mathcal{N}(0, \Sigma_{\dot{x}})$ with diagonal covariance.

The observations are related to the state of a person state by means of a linear transportation matrix $G$ plus Gaussian observation noise $v_k$:

\(^1\)The author would like to acknowledge Dr. Ba-Ngu Vo from the Department of Electrical & Electronic Engineering of the University of Melbourne for the code implementing the OSPA metric.
Figure 5: A single person is represented as an ellipsoid consisting of a 2-dimensional centroid and the equidensity contour of the Gaussian observation noise

\[ G = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \]

\[ y_k = Gx_k + v_k \]

\[ v_k \sim N(0, \Sigma_{v_k}) \]

Figure 5 shows different examples for an elliptical person model.

6.1. Indoor tracking with occlusions

In the first experiment, the indoor_tracking_video_2 sequence from the VISOR dataset\(^2\) is used to illustrate the proposed technique for tracking with occlusions. A temporal Gaussian background model using the parameters specified in Table 1 was used for generating the foreground blobs.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame buffer (frames)</td>
<td>30</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 1: Parameter settings for the background subtraction model

The SMC implementations of the PHD filter and the FB-PHD and TF-PHD smoothers are used to recursively estimate the number of persons and their locations. Parameters for the filter are shown in Table 2 and the cardinality estimates are shown in Figure 6. The PHD filter is not able to correctly estimate the number of persons in the presence of occlusions (frame 287 of the sequence). Because there are no detected persons, the PHD filter estimate is strongly biased to the error, leaving all particles with negligible weights [41]. The FB-PHD smoother is able to alleviate this effect in a backward pass (see Figure 6(a)). However, this is not the case for the TF-PHD smoother which also uses the observations in order to compute the backward estimate.

\(^2\)Video Surveillance Online Repository: http://imagelab.ing.unimore.it/visor/
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of particles per target</td>
<td>1000</td>
</tr>
<tr>
<td>Poisson clutter rate (per unit value)</td>
<td>$5e - 5$</td>
</tr>
<tr>
<td>Poisson birth rate (per unit value)</td>
<td>$1e - 5$</td>
</tr>
<tr>
<td>uniform spatial clutter density</td>
<td>$U([1, 352] 	imes [1, 288])$</td>
</tr>
<tr>
<td>uniform spatial birth density</td>
<td>$U([1, 352] 	imes [1, 288])$</td>
</tr>
<tr>
<td>initial Poisson birth rate</td>
<td>3</td>
</tr>
<tr>
<td>target process noise ($\Sigma_{x_k}$)</td>
<td>diag(15, 15, 3, 3)</td>
</tr>
<tr>
<td>target observation noise ($\Sigma_{y_k}$)</td>
<td>diag(10, 50)</td>
</tr>
<tr>
<td>target survival rate</td>
<td>1</td>
</tr>
<tr>
<td>target detection rate</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table 2: Parameter settings for the PHD filter and smoother.

Figure 6: Cardinality estimates for the PHD filter and smoother. The PHD filter (plotted in dashed lines) fails to estimate the number of persons in the presence of occlusion in frame 287. The FB-PHD smoother is able to recover from the error in a backward pass, but this is not the case for the TF-PHD smoother.

Figure 7 shows the Monte Carlo approximation to the PHD filter for the frame number 67 of the sequence. Location estimates are then obtained by using clustering techniques and the number of clusters corresponds to the PHD cardinality estimates. Both PHD smoothers are able to reduce uncertainty by means of removing spurious samples from the forward pass (see Figures 7(b) and 7(c)).

6.2. People counting and tracking in crowded environments

In this worked exampled, the practical implications of using the PHD filtering in human tracking in real world surveillance scenarios are studied. For that purpose, a benchmark pedestrian database is used which is publicly available for testing new algorithms in crowd analysis. The UCSDPEDES³ dataset contains several videos of pedestrians taken from a stationary surveillance camera. The videos are 8-bit gray

³http://www.svcl.ucsd.edu/projects/peoplecnt/
scale, with dimensions $[238 \times 158]$ at 10 frames per second. We focus on the persons counting and tracking task, and the worked examples will show the PHD performance for this case. Figure 8 shows an example of a particular scene from the dataset.

Multiple observations from a single person caused by over-segmentation would cause problems in multi-target tracking methods. Moreover, incorrect person detections would worsen the SNR ratio, deteriorating the performance of the filter. In Figure 9 (frame 20 of the vidf1_33_001.y sequence of the dataset), the ellipses are used to enclose detected persons and due to the under-segmentation problem, a group of pedestrians is represented by a single target. Furthermore, because no person recognition has been performed, the estimates are not sensitive to the area occupied by a single person. Therefore, as a consequence of a poor SNR ratio, cardinality and state estimates becomes susceptible to under-segmentation and over-segmentation issues. Also, since the PHD filter does not perform any data association, the assessment of the error on individual person locations and velocities is not straightforward, requiring an additional step. Parameters for the PHD filter and smoothers are shown in Table 3.

A person with a bicycle has a bigger area than the expected average, and as a result over-segmentation causes the PHD filter in Figure 9(a) to incorrectly estimate the number of targets in that area. Nevertheless, the TF-PHD smoother in Figure 9(b) is able to give an improved estimate in the region containing a single person.

The estimated number of targets in the backward step is less sensitive to fluctuations in the number of observations (see Table 4). Since estimates and ground truth might have different cardinalities, the OSPA error is used for comparison purposes. Figure 10 shows the estimated number of persons for the PHD filter and both smoothers for the first 50 frames of the sequence.
Figure 8: Crowded scenario with multiple people walking in different directions. A single camera captures images at 10 frames per second and the goal is to track and count individual persons.

Figure 9: Particle PHD filter and TF-PHD smoother estimates for frame 20.

Figure 10: Crowd counting estimates using the PHD filtering and smoothing. Both, the TF-PHD and the FB-PHD smoothers give an improved estimate of the number of targets.
### Table 3: Parameter settings for the PHD filter and smoother.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of particles per target</td>
<td>150</td>
</tr>
<tr>
<td>Poisson clutter rate (per unit value)</td>
<td>$1e^{-4}$</td>
</tr>
<tr>
<td>Poisson birth rate (per unit value)</td>
<td>$1e^{-5}$</td>
</tr>
<tr>
<td>uniform spatial clutter density</td>
<td>$U([1, 238] \times [1, 152])$</td>
</tr>
<tr>
<td>uniform spatial birth density</td>
<td>$U([1, 238] \times [1, 152])$</td>
</tr>
<tr>
<td>initial Poisson birth rate</td>
<td>10</td>
</tr>
<tr>
<td>target process noise ($\Sigma_{x_k}$)</td>
<td>$diag([5.5, 1.1])$</td>
</tr>
<tr>
<td>target observation noise ($\Sigma_{y_k}$)</td>
<td>$diag([8.4])$</td>
</tr>
<tr>
<td>target survival rate</td>
<td>0.95</td>
</tr>
<tr>
<td>target detection rate</td>
<td>0.95</td>
</tr>
</tbody>
</table>

### Table 4: OSPA error (with parameters $p=2, c=2$) for the PHD filter and fixed-interval smoothing for visual tracking.

<table>
<thead>
<tr>
<th>Error</th>
<th>PHD filter</th>
<th>FB-PHD</th>
<th>TF-PHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMS</td>
<td>2.23</td>
<td>1.62</td>
<td>1.53</td>
</tr>
<tr>
<td>OSPA (EM)</td>
<td>1.61</td>
<td>1.61</td>
<td>1.60</td>
</tr>
<tr>
<td>OSPA (Gibbs)</td>
<td>1.61</td>
<td>1.62</td>
<td>1.61</td>
</tr>
</tbody>
</table>

Person locations that are incorrectly addressed due to cardinality errors (wrongly estimated number of persons in the crowd) in the forward pass can be re-estimated in a backward pass. However, since re-sampling was performed in both steps, it is more challenging for the PHD smoothers to provide improved location estimates. Furthermore, since the PHD filter proposes individual samples for each person, location estimates are not sensitive to inter person distances. This issue is also inherited by particle PHD smoothers, so location estimates suffer from the same problem. Figure 11 shows the PHD filter, the FB-PHD and the TF-PHD smoothers using the EM algorithm and the Gibbs sampler in frame 14 of the dataset.
Figure 11: Particle approximations for frame 14 of the pedestrian tracking sequence. Location estimates from the PHD filter suffer from an incorrectly estimated number of persons. Since the Gibbs sampler is less sensitive to the initial conditions, it manages to allocate person locations more accurately and with less variance than the EM algorithm. Monte Carlo approximations by means of the FB-PHD and the TF-PHD smoothers provide improved estimates over the PHD filter alone.
Now the performance of the PHD smoothing approach on the sequence vidf1_33_001.y using fixed-lag implementations is analyzed. As opposed to fixed-interval, fixed-lag implementations can be implemented in real time using a small time lag. Four different time lags are considered and Table 5 shows the performance of the TF-PHD and the FB-PHD smoothers when the EM algorithm and the Gibbs sampler are used for state estimation. In this case we expected to have a large number of outliers in the estimated locations. Therefore, in order to measure the performance of smoothing over filtering, we choose the OSPA metric to be less sensitive to outliers.

<table>
<thead>
<tr>
<th>Error</th>
<th>PHD</th>
<th>FB-PHD</th>
<th>TF-PHD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed-lag (1 time step) implementation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMS</td>
<td>2.26</td>
<td>2.11</td>
<td>2.11</td>
</tr>
<tr>
<td>OSPA (EM)</td>
<td>1.60</td>
<td>1.61</td>
<td>1.60</td>
</tr>
<tr>
<td>OSPA (Bayes)</td>
<td>1.62</td>
<td>1.60</td>
<td>1.60</td>
</tr>
<tr>
<td><strong>Fixed-lag (2 time steps) implementation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMS</td>
<td>2.26</td>
<td>2.04</td>
<td>2.02</td>
</tr>
<tr>
<td>OSPA (EM)</td>
<td>1.60</td>
<td>1.59</td>
<td>1.60</td>
</tr>
<tr>
<td>OSPA (Bayes)</td>
<td>1.62</td>
<td>1.59</td>
<td>1.60</td>
</tr>
<tr>
<td><strong>Fixed-lag (3 time steps) implementation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMS</td>
<td>2.26</td>
<td>1.88</td>
<td>1.86</td>
</tr>
<tr>
<td>OSPA (EM)</td>
<td>1.60</td>
<td>1.57</td>
<td>1.57</td>
</tr>
<tr>
<td>OSPA (Bayes)</td>
<td>1.62</td>
<td>1.58</td>
<td>1.59</td>
</tr>
<tr>
<td><strong>Fixed-lag (5 time steps) implementation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMS</td>
<td>2.26</td>
<td>1.83</td>
<td>1.81</td>
</tr>
<tr>
<td>OSPA (EM)</td>
<td>1.60</td>
<td>1.57</td>
<td>1.57</td>
</tr>
<tr>
<td>OSPA (Bayes)</td>
<td>1.62</td>
<td>1.57</td>
<td>1.57</td>
</tr>
</tbody>
</table>

Table 5: Cardinality and OSPA \((c=2,p=2)\) error for the PHD filter and smoothers for visual tracking

Increasing the time-lag improves performance, but it can be seen that the OSPA error for both EM and Gibbs sampler estimation converges at time lag 5.

7. Conclusion

An important remark on PHD filter for visual tracking can be discussed in terms of whether measurement-to-measurement and measurement-to-track associations are available or not. If a particular tracking scenario in consideration allows us to concatenate multiple single target filters, then standard multi-hypothesis approach will perform seamlessly without any distributional assumption (e.g. first-order moment approximations). However, if we cannot override clutter using gating techniques or we cannot distinguish between a new-born or an existing target, the algorithm would potentially end up having a combinatorial explosion in the number of association hypotheses.

The PHD filter was originally conceived in a somehow different scenario, where the expected value of the unknown number of targets is calculated by estimating the ratio of false measurements and the likelihood of a single target. Such modeling is
useful in highly cluttered environments with targets having large signal-to-noise ratio. This setup is not always well suited in visual tracking, where the first-order moment approximation has an adversarial effect in the estimation procedure which cannot always be alleviated in a backward pass. Nevertheless, we demonstrated the benefits of two PHD smoothing techniques for estimating persons locations.

Further work will consider integrating person detection schemes into the PHD filter. Using this approach, the likelihood of a single person would not only consider the false alarms ratio but also the geometry or the shape of each person being also defined by random parameters. Moreover, this stochastic model would also allow to depart from the first-order moment approximation to the posterior, including persons interactions and larger occlusions.

References


22


