Occlusion handling for online visual tracking using labeled random set filters

Rathnayake, Tharindu; Khodadadian Gostar, Amirali; Hoseinnezhad, Reza; Bab-Hadiashar, Alireza

Rathnayake, T., Khodadadian Gostar, A., Hoseinnezhad, R., & Bab-Hadiashar, A. (2017). Occlusion handling for online visual tracking using labeled random set filters. Proceedings of the 6th International Conference on Control, Automation and Information Sciences (ICCAIS 2017), 151–156. https://doi.org/10.1109/ICCAIS.2017.8217567
Document Version: Accepted Manuscript

Published Version: https://doi.org/10.1109/ICCAIS.2017.8217567

Occlusion Handling for Online Visual Tracking Using Labeled Random Set Filters

Tharindu Rathnayake Amirali Khodadadian Gostar Reza Hoseinnezhad Alireza Bab-Hadiashar
School of Engineering
RMIT University, Victoria 3083, Australia
Emails: {tharindu.rathnayake, amirali.khodadadian, rezah, abh}@rmit.edu.au

***Abstract*—This paper presents a novel solution to the occlusion handling problem in pedestrian tracking using labeled random finite set theory. The occlusion handling module uses motion and color cues of tracked targets to recover target labels after occlusion. An effective algorithm is also proposed for false alarm detection and removal which is designed based on tracked targets features such as, *overlap ratio*, *size similarity* and the *time of track initialization* of the tracked targets. We implement our solution using sequential Monte Carlo method, and compare it with state-of-the-art visual tracking methods. The results show that the proposed algorithm perform favorably in terms of various standard performance metrics.**

I. Introduction

Online visual multi-target tracking is one of the challenging and ubiquitously addressed problems in computer vision. These techniques use the detections in the current and previous frames to estimate the states of the targets at each time step [1]–[4] which is in stark contrast to the batch processing methods. Batch processing methods utilize the extracted information from the entire sequence of frames and iteratively optimize the detection assignment of the current frame using past and future information [5]–[9]. While they manage miss-detections better than online methods [10], they can not be used in *real-time* applications, such as surveillance.

One of the prominent challenges in visual multi-target tracking is long and short term occlusions. The solutions implemented for this issue include using occlusion geodesics [11], connecting short tracklets to form longer trajectories and bridging the gaps due to occlusions [12], and considering relative overlap, depth ordering and visibility of targets to formulate analytical global occlusion models [6].

In this paper we propose a new solution for online occlusion handling. We adapt the Vo-Vo filter for online visual multi-target tracking. Vo-Vo filter or δ-Generalized Labeled Multi-Bernoulli, (δ-GLMB) filter is a Bayesian recursion filter which has been applied successfully in radar tracking [13], [14]. We incorporate the targets' motion information into the state space as well as the detection information in the form of rectangular boxes. Further, we propose a region of interest based birth process model for handling the initialization of new target trajectories as well as re-detecting missing targets.

A novel merging procedure is also proposed and implemented to avoid one target being represented by multiple blobs. In order to handle long term occlusion events, we introduce a novel track management algorithm which is henceforth referred to as "label recovery procedure". In formulating the label recovery procedure, various aspects are considered such as the number of time steps between the disappearance and re-detection of the target, the features of the disappeared and re-detected targets and the spatial distance between them.

The rest of the paper is organized as follows. Section II briefly reviews the foundations of RFS theory and the notation used, followed by a problem statement. Then section III explains Vo-Vo filter as the tracking algorithm used in our proposed method. We then present our solution for occlusion recovery in section IV. Section V evaluates the proposed method on publicly available datasets with comparative results, followed by concluding remarks presented in section VI.

II. Background and Problem Statement

In order to formally state the problem of occlusion in visual tracking, we first introduce notation and background in stochastic multi-target filtering and the Bayesian recursion that is used in implementation of such filters.

A. Notation and background

The ensemble of multiple targets and their labels is denoted by $\boldsymbol{X} = \{(x, \ell)\}$ where $x \in \mathbb{X}$ is a target state and $\mathbb{X}$ is the state space and $\ell \in \mathbb{L}$ is the label associated with target state x and $\mathbb{L}$ is the label space. In visual tracking literature, a common choice for target state is formulated based on treating targets as rectangular blobs. In that case, the state includes the location and dimensions of the target blob, and perhaps their time-derivatives. For example, in our experiments, each target state is a 6-tuple in the form of

$$x = \begin{bmatrix} p_x & p_y & \dot{p_x} & \dot{p_y} & w & h \end{bmatrix}^\top,$$

where p_x and p_y denote the image coordinates of the center of the blob (target location), $\dot{p_x}$ and $\dot{p_y}$ are the velocities in x and y image coordinate directions and w and h denote the width and height of the blob, respectively.

In the random set multi-target tracking literature, the label is usually defined as a pair $\ell = (k_b, i_b)$ where k_b is the time step k at which the target is *born* (enters the scene) and i_b is an *index* to distinguish different targets born at the same time step.

A labeled RFS is a labeled finite set that admits random variations both in its number of elements (its cardinality), values of the elements and labels. In stochastic filtering, those

random variations are modeled by a statistical density denoted by $\boldsymbol{\pi}(\boldsymbol{X})$ whose parameters are recursively determined in a *Bayesian recursion* scheme using measurements (detections) and stochastic models for targets' motions and their birth and death.

As for the Bayesian recursion, the procedure can be split into two steps: prediction and update. Stochastic models for target's state evolution, birth and death are implemented in the *prediction* step. Detections which are acquired from sensor(s) are then used in the *update* step. Let the set of observations (detections) at time k be denoted by Z_k, and all the observations acquired up to time k be denoted by $Z_{1:k}$. Bayesian recursion transforms a prior multi-target density $\boldsymbol{\pi}_k(\boldsymbol{X}|Z_{1:k})$ to a posterior multi-target density $\boldsymbol{\pi}_{k+1}(\boldsymbol{X}|Z_{1:k+1})$ from which the number and states (including labels) of existing targets can be inferred. Indeed, the labeled multi-target density is recursively predicted (based on Chapman-Kolmogorov equation) and updated (using Bayes' rule) using [13]:

$$\boldsymbol{\pi}_{k+1|k}(\boldsymbol{X}|Z_{1:k}) = \int \boldsymbol{f}_{k+1|k}(\boldsymbol{X}|\boldsymbol{X}_k)\boldsymbol{\pi}_k(\boldsymbol{X}_k|Z_{1:k})\delta\boldsymbol{X}_k \quad (1)$$

$$\boldsymbol{\pi}_{k+1}(\boldsymbol{X}|Z_{1:k+1}) = \frac{g_{k+1}(Z_{k+1}|\boldsymbol{X})\boldsymbol{\pi}_{k+1|k}(\boldsymbol{X})}{\int g_{k+1}(Z_{k+1}|\boldsymbol{X}_k)\boldsymbol{\pi}_{k+1|k}(\boldsymbol{X}_k)\delta\boldsymbol{X}_k} \quad (2)$$

where $\boldsymbol{f}_{k+1|k}(\cdot|\cdot)$ is the multi-target transition density from time k to $k+1$, $g_{k+1}(Z_{k+1}|\boldsymbol{X})$ is the multi-target likelihood of the measurement set Z_{k+1} conditioned on labeled multi-target state $\boldsymbol{X}$, and the integrals are set integrals defined as:

$$\int f(\boldsymbol{X})\delta\boldsymbol{X} = \sum_{i=0}^{\infty}\sum_{(\ell_1,\ell_2,...,\ell_i)\in\mathbb{L}^i} \int_{\mathbb{X}^i} f((x_1,\ell_1),...,(x_i,\ell_i))d(x_1,...,x_i).$$

B. Problem statement

The multi-target posterior density $\boldsymbol{\pi}_{k+1}(\cdot|\cdot)$ captures all information on the number of targets and their individual states. The multi-target likelihood function $g_{k+1}(\cdot|\cdot)$ encapsulates information about target models and false alarm models. The multi-target transition density $\boldsymbol{f}_{k+1|k}(\cdot|\cdot)$ describes the motion, birth and death of targets.

At any time k, the multi-target posterior density $\boldsymbol{\pi}(\boldsymbol{X})$ can be directly used to estimate the number of targets $\hat{n}_k$ and their labels and states

$$\hat{\boldsymbol{X}}_k = \{(\hat{x}_{i,k},\hat{\ell}_{i,k})\}_{i=1}^{\hat{n}_k}. \quad (3)$$

Consider a target born at time k_b with index i_b, so its label is $\ell = (k_b, i_b)$. At a later time step $k > k_b$, assume that the tracking algorithm performs well enough to include this label within its labeled set estimate $\hat{\boldsymbol{X}}_k$. Suppose that at time k, this target is occluded by another, and for a period of time, T_o, the target disappears from the measurement sets $Z_{(k+1):(k+T_o)}$. As a result of this, the target label will disappear from the estimates returned by the multi-target filter with a delay (the delay in disappearance depends on filter parameters). Thus, if the occlusion period T_o is not too short, after occlusion (when the target with label ℓ reappears among the detections) it will be treated as a new target with a different label $\ell' = (k'_b, i'_b)$ where $k'_b = k + T_o + 1$. The problem is to distinguish such an occluded then reappearing target from a newly born target, and recover its label.

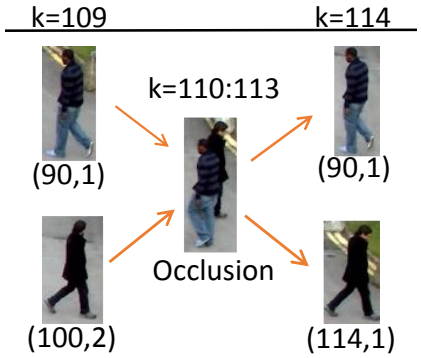

Fig. 1. An example demonstrating label ambiguity after an occlusion, and the need for label recovery.

To further clarify the problem, an example is shown in Figure 1. The pedestrians with labels (90,1) and (100,2) are correctly tracked until time step 110, at which point one target occludes the other. Depending on the frame rate of the imaging device, the occlusion period may take several time steps or frames (in this example, until time step 113) during which, the target detection module may return only one measurement for both pedestrians. After the occlusion, both pedestrians reappear at time $k = 114$, and the Bayesian multi-target filtering algorithm starts tracking the previously occluded pedestrian by assigning a new label (114,1). This is not acceptable in many visual tracking applications such as security and surveillance as the occluded target should be re-identified and not treated as a new person.

Remark 1. *The target birth model that is incorporated in the multi-target transition density, is usually designed based on the prior knowledge that we have about the regions of entry for new targets. In the explanation given above for label recovery problem as part of an occlusion handling solution, we assume that birth region is expanded to include the whole surveillance region. Only with this expansion we make sure that re-detected targets after an occlusion appear among the estimates with the label of a newly born target. If the birth region is kept limited (e.g. only around the edges of the camera field of view), then a target that is occluded in the middle of the field of view, can forever disappear from the estimates returned by the Bayes filter even if it appears among the detections after occlusion.*

Remark 2. *Expansion of the birth region can cause measurements that are associated with existing targets to strengthen wrong birth hypotheses in the vicinity of those targets. Hence, the resulting estimate may include non-existing target estimates around existing targets. A complete occlusion handling solution must include a remedy for detection and removal of such false positives (false alarms).*

III. Labeled Random Set Filtering

The Vo-Vo filter is formulated based on propagating a Vo-Vo density through Bayesian recursion. A Vo-Vo density on

$\mathbb{X} \times \mathbb{L}$ is defined as [15]:

$$\boldsymbol{\pi}(\mathbf{X}) = \Delta(\mathbf{X}) \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} w^{(I,\xi)}\delta_I(\mathcal{L}(\mathbf{X}))[p^{(\xi)}]^{\mathbf{X}} \quad (4)$$

where I denotes a set of track labels and ξ denotes a realization of a discrete space Ξ which represents the history of track labels to measurement associations. This distribution can be interpreted as a weighted mixture of exponentials of multi-target densities. The function $\Delta(\boldsymbol{X}) \triangleq \delta_{|X|}(|\mathcal{L}(\boldsymbol{X})|)$ returns 1 only if different targets in $\boldsymbol{X}$ are assigned different labels, otherwise it returns zero. The weights $w^{(I,\xi)}$ and the spatial distributions $p^{(\xi)}$ satisfy the normalization conditions

$$\sum_{(I,\xi)\in\mathcal{F}(\mathbb{L})\times\Xi} w^{(I,\xi)} = 1, \ \int p^{(\xi)}(x,\ell)dx = 1. \quad (5)$$

With the standard multi-target model, the Vo-Vo density is closed under the Chapman-Kolmogorov equation and also a conjugate prior with standard multi-target likelihood (i.e. both predicted and updated densities are Vo-Vo densities in the form of (4)) [13], [15].

Indeed, the prediction step leads to the following Vo-Vo density: [13]

$$\boldsymbol{\pi}(\mathbf{X}) = \Delta(\mathbf{X}) \sum_{(I,\xi)\in\mathcal{F}(\mathbb{L}_+)\times\Xi} w_+^{(I,\xi)}\delta_I(\mathcal{L}(\mathbf{X}))[p_+^{(\xi)}]^{\mathbf{X}} \quad (6)$$

where:

$$\begin{aligned}
w_+^{(I,\xi)} &= w_S^{(\xi)}(I\cap\mathbb{L})\ w_B(I\cap\mathbb{B}) \\
p_+^{(\xi)}(x,\ell) &= 1_{\mathbb{L}}(\ell)p_S^{(\xi)}(x,\ell) + 1_{\mathbb{B}}(\ell)p_B(x,\ell) \\
w_S^{(\epsilon)}(L) &= [\eta_S^{(\xi)}]^L \sum_{I\supseteq L}[1-\eta_S^{(\xi)}]^{I-L} w^{(I,\xi)} \\
\eta_S^{(\xi)}(\ell) &= \langle p_S(\cdot,\ell), p^{(\xi)}(\cdot,\ell)\rangle \\
p_S^{(\xi)}(x,\ell) &= \frac{\langle p_S(\cdot,\ell)f(x|\cdot,\ell), p^{(\xi)}(\cdot,\ell)\rangle}{\eta_S^{(\xi)}(\ell)}
\end{aligned}$$

and $p_S(x,\ell)$ denotes the probability of survival for a labeled target with state $\mathbf{x} = (x,\ell)$, $f(x|\cdot,\ell)$ denotes the single-target state transition density, and $w_B(I)$ and $p_B(x,\ell)$ are the parameters of the following labeled birth density defined as a special case of Vo-Vo density on the birth space $\mathbb{B}$ as follows: [13]

$$\boldsymbol{\pi}_B(\mathbf{X}) = \Delta(\mathbf{X})w_B(\mathbf{X})[p_B]^{\mathbf{X}}. \quad (7)$$

The label space is also extended to include the newly born targets, $\mathbb{L}_+ = \mathbb{L}\cup\mathbb{B}$.

The updated Vo-Vo density is also of the mathematical form given in (4), denoted by:

$$\begin{aligned}
\boldsymbol{\pi}(\mathbf{X}|Z) = \Delta(\mathbf{X}) \textstyle\sum_{I\in\mathcal{F}(\mathbb{L}_+)} \sum_{\xi\in\Xi} \sum_{\theta\in\Theta(I)} \\
w^{(I,\xi,\theta)}(Z)\delta_I(\mathcal{L}(\mathbf{X}))\left[p^{(\xi,\theta)}(\cdot|Z)\right]^{\mathbf{X}}
\end{aligned} \quad (8)$$

where $\Theta(I)$ is the subset of current association maps from the label set I to the measurement set Z in the sense that the target labeled $\ell \in I$ is associated with measurement $z_{\theta(\ell)}$. By convention, for the missed targets, $\theta(\ell) = 0$. According to [13] the weights and densities of the updated components of the Vo-Vo density are given by:

$$\begin{aligned}
w^{(I,\xi,\theta)}(Z) &\propto w_+^{(I,\xi)}[\eta_Z^{(\xi,\theta)}]^I \\
p^{(\xi,\theta)}(\cdot|Z) &= p_+^{(\xi)}(x,\ell)\ \psi_Z(x,\ell;\theta)/\eta_Z^{(\xi,\theta)}(\ell)
\end{aligned}$$

where:

$$\begin{aligned}
\eta_Z^{(\xi,\theta)}(\ell) &= \langle p_+^{(\xi)}(\cdot,\ell), \psi_Z(\cdot,\ell;\theta)\rangle \\
\psi_Z(x,\ell;\theta) &= \begin{cases} 1-p_D(x,\ell) & \text{if } \theta(\ell)=0 \\ p_D(x,\ell)g(z_{\theta(\ell)}|x,\ell)/\kappa(z_{\theta(\ell)}) & \text{otherwise} \end{cases}
\end{aligned}$$

where $g(z|x,\ell)$, $p_D(x,\ell)$ and $\kappa(z)$ are the single-target likelihood, detection probability and clutter intensity function, respectively.

Remark 3. *To avoid exponential explosion of the number of hypotheses, those with very small weights (less than 10^{-5} in this work) are pruned, and the weights of the remaining hypotheses are renormalized. Furthermore, the maximum number of allowed hypotheses is 700. In a sequential Monte Carlo (SMC) implementation, the particles representing the single target densities $p^{(\xi,\theta)}$ are also need to be resampled. A combination of ranked assignment and shortest-path strategies is also suggested in [13] for computationally efficient implementation of the prediction and update steps.*

Remark 4. *Given a posterior Vo-Vo density in the form of* (8)*, the discrete distribution of number of targets (cardinality) is given by:*

$$\rho(n) = \sum_{\substack{I\in\mathcal{F}(\mathbb{L}_+) \\ |I|=n}} \sum_{\xi\in\Xi}\sum_{\theta\in\Theta(I)} w^{(I,\xi,\theta)}(Z) \quad (9)$$

and a maximum a posteriori (MAP) estimate for the number of targets is:

$$\hat{n} = \arg\max_n \ \rho(n). \quad (10)$$

Defining:

$$(I^*,\xi^*,\theta^*) = \underset{\substack{I\in\mathcal{F}(\mathbb{L}_+),\ |I|=\hat{n} \\ \xi\in\Xi,\ \theta\in\Theta(I)}}{\arg\max} w^{(I,\xi,\theta)}(Z), \quad (11)$$

the estimated set of labeled target states is:

$$\hat{\mathbf{X}} = \{(\hat{x}(\ell),\ell)\}_{\ell\in I^*} \quad (12)$$

with $\hat{x}(\ell) = \int x\ p^{(\xi^*,\theta^*)}(x,\ell|Z)\ dx$.

IV. Proposed Method

To handle occlusions in multi-target visual tracking, we suggest a combination of false alarm detection and removal as well as label recovery process that operates on the labeled set estimate returned by the Vo-Vo filter. Our proposed method only needs to compute a limited number of mutual distances and memorize a few color histograms, because it only operates on estimates and not the entire ensemble of labeled multi-target hypotheses in the Vo-Vo posterior. In the following sections we separately address two problems: A. the false alarm detection and removal B. label recovery methods. We note that in principle, the mentioned operations would be formulated for

implementation within the update step of the filter. However, such implementations would involve computation of a large number of mutual distances and storing a substantial number of color histograms, thus being computationally expensive for an online visual tracking system.

A. False alarm removal and detection

Consider the multi-target estimate $\hat{\mathbf{X}}$ returned by the Vo-Vo filter. The algorithm compares each target with all the other targets in the labeled set for false alarms. For a labeled target $\boldsymbol{y}' = (x^{(\ell')}, \ell') \in \hat{\mathbf{X}}$ to be detected as false alarm and removed from the estimate, it should satisfy the following three conditions in terms of its similarities with another detected target $\boldsymbol{y} = (x^{(\ell)}, \ell) \in \hat{\mathbf{X}}$:

1) The two targets must have substantial overlap.
2) $\boldsymbol{y}$ must be older than $\boldsymbol{y}'$.
3) The two targets must be of similar size.

The rationale behind the first two conditions is that we are looking for false alarms that are caused by birth targets that match measurements, that are already covered by existing targets. Hence, false alarms are expected to significantly overlap existing targets. Furthermore, being the result of birth process, the false alarms are expected to have been born after the real targets with which they have substantial overlap.

The algorithm searches for all pairs of targets in the estimate $\hat{\mathbf{X}}$ that substantially overlap, and removes the label with new time stamp as a false alarm. There might be two real targets, one far from and the other close to the camera, and the closer target (larger in the image) may cover a substantial portion of the farther one. In this case, both targets are real, and no false alarm should be detected and removed. This is the main rationale behind the third condition.

Having the time of birth recorded as part of the target's label, makes it straightforward to distinguish which of two targets is older. If $\mathcal{L}(\boldsymbol{y}) = (k_b, i_b)$ and $\mathcal{L}(\boldsymbol{y}') = (k_b', i_b')$, then we have:

$$\textsc{Older}(\boldsymbol{y}, \boldsymbol{y}') = \begin{cases} \boldsymbol{y} & \text{if } k_b < k_b' \\ \boldsymbol{y}'. & \text{otherwise} \end{cases} \tag{13}$$

Upon finding each false alarm, we remove it from the track table of the filter so that it does not propagate into the next time step.

B. Label recovery

As it was mentioned earlier, in many visual tracking applications, either due to the shortcomings of the employed detector or occlusion, targets may not be tracked and they can temporarily disappear from the trajectories returned by the filter. When a target is re-detected (e.g. after occlusion), the filter can include the target in its estimate but as a new trajectory (with a new label). In some tracking applications such as surveillance, it is of paramount importance that the targets have consistent labels before and after such temporary disappearances. Inspired by the decay functions in distance dependent Chinese restaurant processes [16], we propose a novel label recovery module to consistently maintain the labels of the targets in occlusion and miss-detection events.

Our proposed label recovery solution is based on constructing a *recent disappearance lookup table* that holds all the targets that have disappeared during the past $k_{\max}$ time steps and have not reappeared yet. The parameter $k_{\max}$ is practically the maximum duration of occlusion that is expected to be handled by our method. The lookup table is constructed as follows.

Let us denote the multi-target estimate returned by the filter at time k by $\hat{\mathbf{X}}_k$. For every single-target state $\mathbf{x} \in \hat{\mathbf{X}}_{k-1}$, it is considered as disappeared at time k if its label does not appear in the set of estimated labels at time k, i.e. if $\mathcal{L}(\mathbf{x}) \notin \mathcal{L}(\hat{\mathbf{X}}_k)$. In that case, the time of disappearance, k, the label of the target $\mathcal{L}(\mathbf{x}) = (k_b, i_b)$, its location (p_x, p_y) and the color histogram of the contents of the target in the image, denoted by H, are all stored in the lookup table. This means appending a new row at the bottom of the lookup table, with contents $\begin{bmatrix} k & k_b & i_b & p_x & p_y & H \end{bmatrix}$. To constrain its size, at any time k, all the recorded rows with birth time labels $k_b < k - k_{\max}$ are removed.

For label recovery, we first find the set of all the newly born targets at time k among the estimates returned in $\hat{\mathbf{X}}_k$,

$$\hat{\mathbf{X}}_{B,k} = \left\{ \mathbf{x} \in \hat{\mathbf{X}}_k \mid \exists i_b \in \mathbb{N}; \mathcal{L}(\mathbf{x}) = (k, i_b) \right\}. \tag{14}$$

For each newly born target estimate $\mathbf{x}$, we then evaluate its similarity to each of the previously disappeared targets recorded in the lookup table. Let us assume that (p_x, p_y) is the location of $\mathbf{x}$, and H is its color content histogram. Consider a previously disappeared target that is recorded in the j-th row of the lookup table as $\begin{bmatrix} k_j & k_{j,b} & i_{j,b} & p_{x_j} & p_{y_j} & H_j \end{bmatrix}$. We are interested in an intuitive and effective technique to quantify the likelihood of $\mathbf{x}$ representing the reappearance of the above mentioned j-th recorded target. Hereafter, we denote this likelihood by $l_j(\mathbf{x})$ and is formulated based on the following intuitions.

In visual tracking applications, one would intuitively expect a disappearing target to maintain its visual appearance (hence its color content histogram) when reappearing. The similarity in visual appearance can be quantified in terms of the distance between the two color histograms. A common choice for formulating such a distance is the Bhattacharyya distance [2], [17].

In addition to similarities in color contents, depending on the period of disappearance, there would be a constrained area in which the target can possibly reappear. Considering the most general model, the random walk, such an area is a disk around (p_{x_j}, p_{y_j}), with a diameter that is proportional to the hypothesized period of disappearance, $k - k_j$.

Based on the above constraints, we suggest to quantify the likelihood of $\mathbf{x}$ representing the j-th recorded disappearance in the lookup table, as follows:

$$\begin{aligned} l_i(\mathbf{x}) \quad \propto \quad & \beta \ \exp\left(-\frac{\sqrt{(p_x - p_{x_j})^2 + (p_y - p_{y_j})^2}}{2[(k - k_j)\sigma_v]^2} \right) \\ & + (1 - \beta) \ \exp\left(-\frac{d(H, H_j)^2}{2\sigma_H^2} \right) \end{aligned} \tag{15}$$

where $d(H, H')$ denotes the Bhattacharyya distance between the two histograms, $\beta \in [0, 1]$ is the weight given to spatial component of the likelihood function, σ_v is the scale of noise in random walk motion model in pixels, and σ_H is the standard deviation of possible random changes in a target's appearance (its color content histogram) from one frame to another. Note that the weighted sum in the right hand side of equation (15) is normalized.

The optimal choice of β parameter depends on the application. For example, if there is no appearance information or all targets of interest have similar appearances, lower emphasis on the appearance component and more on the spatial component (larger β) is suitable. In cases where the targets can be easily distinguished from their color features, one can assign a larger weight for the appearance component (smaller β).

For each element $\mathbf{x}$ in the newly born estimates, its likelihood to be a reappearance of all the previously disappeared targets is computed, and the best candidate (with the maximum likelihood) is chosen. If its likelihood is larger than a user-defined threshold l_{th}, it is accepted as a reappearance, and its label is recovered. Noting that the likelihood values in (15) are all normalized to fall within $[0, 1]$, the same is correct for the threshold l_{th}, which was set to 0.7 in our experiments.

V. Experimental Results

In extensive experiments using publicly available datasets, we examined the performance of our visual tracking solution and compared it with the following state-of-the-art methods in the computer vision literature: RMOT [18], StruckMOT [19], GeodesicTracker [11], PRIMPT [20] and Non-linear motion [8].

Due to the perspective effect, the target sizes vary when they move towards or away from the camera. Thus the width and the height of the target states are set to have variable, but constrained values. The targets are set to have a constant survival probability of $p_{S_k}(\cdot) = 0.99$ and are assumed to follow the nearly constant velocity model. In all case studies, the birth processes are labeled multi-Bernoulli with constant probabilities of existence of 0.03. In order to strike the right balance between accuracy of particle approximation and computation, the number of particles per target is constrained between $L_{\text{min}} = 100$ and $L_{\text{max}} = 500$.

In order to permit a fair comparison we use the same set of metrics proposed by Li et al. [21], as those have been widely used in the visual tracking literature [8], [9], [18], [20], [22]. Note that most of the methods used in comparison have used the same detection results and ground truth available in the website[1] of one of the authors of [8], [9], along with the evaluation software. We selected three publicly available datasets which are widely used in the literature [1], [8], [9], [18]–[20], [22], [23] to benchmark the performance of visual tracking algorithms.

[1]http://iris.usc.edu/people/yangbo/downloads.html

1) PETS2009 S2L1 View1: From the comparative results presented in table I, it can be seen that our online method returns generally better values for precision, MT, ML, Frag and IDS metrics when compared to online methods. It should be noted that, although StructMOT reports better results for REC and Frag metrics, it should be trained offline and is not directly comparable online methods such as ours.

2) TUD-Stadtmitte: In this sequence most of the pedestrians have almost similar color features resulting in similar color histograms. Thus motion information is more important than color information in tracking. We assigned a large value for the weight of motion information, β parameter, in our occlusion label recovery algorithm. It also shows that there has been no ID switches which demonstrate excellent label management performance. In addition, this sequence comprised of multiple detections for the same target. The ability to handle clutter in Vo-Vo filter mitigates the effects of these multiple detections.

3) ETH BAHNHOF and SUNNYDAY: The metric values for these sequences are lower compared to that of the other two sequences as there is a large number of occluded targets and number of miss-detections (specially in the SUNNYDAY sequence). In both sequences, when a reflection of a pedestrian appear on the glass, it is detected by the detector and thus tracked by our method, resulting in lower metric values. Furthermore, frequent miss-detections make the fragmentation metric higher.

Due to the space constraints snapshots of frames with tracking results are included as supplemental material.

A. Computation Speed

The algorithm is implemented in MATLAB R2015a in a core i7 laptop with 8GB of memory and the implementation is not optimized. With the particle count mentioned in section V and LMB birth processes mentioned in sections for each dataset, the algorithm is capable of achieving a speed of 4 frames per second for each of the sequences permitting it to be used in real time applications.

VI. Conclusion

In this paper, we presented a novel online visual tracking algorithm with false alarm removal and occlusion handling modules based on RFS theory. The algorithm was implemented using SMC techniques and it was evaluated on a number of standard datasets. A comparison with some state-of-the-art algorithms showed that our method outperformed or comparable to the methods in comparison.

VII. Acknowledgment

This work was supported by ARC Discovery Projects grant DP130104404, and ARC Linkage Projects grant LP130100521.

References

[1] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 9, pp. 1820–1833, 2011.

TABLE I
COMPARATIVE RESULTS FOR PETS2009-S2L1V1, TUD-STADTMITTE AND ETH BAHNHOF AND SUNNYDAY

represents the online methods

Dataset	Method	REC	PRE	FAF	GT	MT	PT	ML	Frag	IDS
PETS09-S2L1	GLMB (Ours)	93.3%	96.9%	0.17	19	94.7%	5.3%	0.0%	20	0
	RMOT [18]	95.6%	95.4%	0.05	19	94.7%	5.3%	0.0%	23	1
	StruckMOT [19] (*o.t.*)	97.2%	93.7%	0.38	19	94.7%	5.3%	0.0%	19	4
	PRIMPT [20] (*o.t.*)	89.5%	99.6%	0.02	19	78.9%	21.1%	0.0%	23	1
	Non-linear motion [8] (*s.c.*)	91.8%	99.0%	0.05	19	89.5%	10.5%	0.0%	9	0
TUD - Stadtmitte	GLMB (Ours)	87.1%	97.1%	0.16	10	80.0%	20.0%	0.0%	6	0
	RMOT [18]	87.9%	96.6%	0.19	10	80.0%	20.0%	0.0%	7	6
	StruckMOT [19] (*o.t.*)	87.3%	95.4%	0.25	10	80.0%	20.0%	0.0%	11	0
	PRIMPT [20] (*o.t.*)	81.0%	99.5%	0.028	10	60.0%	30.0%	10.0%	0	1
	OnlineCRF [9]	87.0%	96.7%	0.18	10	70.0%	30.0%	0.0%	1	0
ETH BAHNHOF and SUNNYDAY	GLMB (Ours)	77.1%	83.6%	1.161	124	54.0%	40.3%	5.6%	91	31
	RMOT [18]	81.5%	86.3%	0.98	124	67.7%	27.4%	4.8%	38	40
	StruckMOT [19] (*o.t.*)	78.4%	84.1%	0.98	124	62.7%	29.6%	7.7%	72	5
	PRIMPT [20] (*o.t.*)	76.8%	86.6%	0.89	125	58.4%	33.6%	8.0%	23	11
	OnlineCRF [9]	79.0%	85.0%	0.64	125	68.0%	24.8%	7.2%	19	11

Note: GT denotes the number of Ground Truth tracks in the case study, *o.t.* represents methods that have to be trained offline and *s.c.* represents the methods that can only be applied to stationary cameras.

[2] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *Computer Vision-ECCV 2004*. Springer, 2004, pp. 28–39.

[3] X. Song, J. Cui, H. Zha, and H. Zhao, *Vision-Based Multiple Interacting Targets Tracking via On-Line Supervised Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 642–655.

[4] V. Takala and M. Pietikainen, "Multi-object tracking using color, texture and motion," in *In CVPR*, 2007.

[5] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1926–1933.

[6] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 58–72, 2014.

[7] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1201–1208.

[8] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1918–1925.

[9] ——, "An online learned CRF model for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2034–2041.

[10] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1815–1821.

[11] H. Possegger, T. Mauthner, P. M. Roth, and H. Bischof, "Occlusion geodesics for online multi-object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1306–1313.

[12] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1200–1207.

[13] B.-N. Vo, B.-T. Vo, and D. Phung, "Labeled Random Finite Sets and the Bayes Multi-Target Tracking Filter," *IEEE TSP*, vol. 62, December 2014.

[14] M. Beard, B.-T. Vo, and B.-N. Vo, "Bayesian multi-target tracking with merged measurements using labelled random finite sets," *IEEE Transactions on Signal Processing*, vol. 63, no. 6, pp. 1433–1447, 2015.

[15] B.-T. Vo and B.-N. Vo, "Labeled random finite sets and multi-object conjugate priors," *Signal Processing, IEEE Transactions on*, vol. 61, no. 13, pp. 3460–3475, 2013.

[16] D. M. Blei and P. I. Frazier, "Distance dependent Chinese restaurant processes," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2461–2488, 2011.

[17] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Computer vision-ECCV 2002*. Springer, 2002, pp. 661–675.

[18] J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon, "Bayesian multi-object tracking using motion context from multiple objects," in *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 33–40.

[19] S. Kim, S. Kwak, J. Feyereisl, and B. Han, "Online multi-target tracking by large margin structured learning," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 98–111.

[20] C.-H. Kuo and R. Nevatia, "How does person identity recognition help multi-person tracking?" in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1217–1224.

[21] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2953–2960.

[22] F. Poiesi, R. Mazzon, and A. Cavallaro, "Multi-target tracking on confidence maps: An application to people tracking," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1257–1272, 2013.

[23] S.-H. Bae and K.-J. Yoon, "Robust online multiobject tracking with data association and track management," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 2820–2833, 2014.