Joint detection and tracking of moving objects using spatio-temporal marked point processes

Paula Craciun, Mathias Ortner, Josiane Zerubia

Joint detection and tracking of moving objects using spatio-temporal marked point processes

Paula Crăciun
INRIA
paula.craciun@inria.fr

Mathias Ortner
Airbus D&S
mathias.ortner@eads.net.fr

Josiane Zerubia
INRIA
josiane.zerubia@inria.fr

Abstract

In this paper, we present a novel approach based on spatio-temporal marked point processes to detect and track moving objects in a batch of high resolution images. Batch processing techniques are applicable to and desirable for a large class of applications such as offline scene and video analysis, and provide better overall detection and data association accuracy than sequential methods. We develop a new, intuitive energy based model consisting of several terms that take into account both the image evidence and physical constraints such as target dynamics, track persistence and mutual exclusion. We construct a suitable optimization scheme that allows us to find strong local minima of the proposed highly non-convex energy. We test our model on three batches of 25 synthetic biological images with different levels of noise. Our main application however consists of two batches of 14 remotely sensed high resolution optical images of boats which are particularly hard to analyze due to the different angles at which the images were taken and the low temporal frequency.

1. Introduction

In a simplistic view, tracking can be defined as the problem of estimating the trajectories of objects in the image plane, as they move around the scene. Hence, a tracker assigns consistent labels to the objects in different frames of a sequence of images. Additionally, it can provide information about the orientation, shape or size of the objects.

Multi-target tracking has been historically achieved using sequential techniques, classical examples of which are the Joint Probabilistic Data Association Filter (JPDAF) or the Multi-Hypothesis Tracker (MHT) [1]. These methods are preferred when real-time object tracking is needed. The major drawback of such methods however is that they cannot modify past results in the light of new data. Nevertheless, real-time tracking is not always a necessary constraint. Applications such as offline video processing or information retrieval do not impose such a constraint. Batch processing methods are preferred in this case since they do not suffer from the limitations of sequential methods. The increased performance of modern hardware allows for new batch processing techniques which could not be explored in the past. In this regard, MCMC Data Association has been proposed in [15] to partition a discrete set of detections into tracks. To retrieve the partitions, a one-to-one target to detection mapping assumption was made. This work was later extended by Yu et al. [21] to overcome the one-to-one mapping assumption. A modified version based on the data association method developed by Yu et al. [21] was applied in video-microscopy [17]. Nevertheless, batch processing techniques remain poorly explored and highly underused.

Marked point processes (MPPs) [18] have been applied with success to object detection problems in high resolution remotely sensed optical images. Applications range from detecting flamingos, buildings or boats [7] but also to people detection in crowds in street view images [10]. The use of specific terms in the energy to be optimized makes these processes application dependent, but highly efficient and accurate. In spite of their good theoretical properties, to our best knowledge marked point processes have never been applied to tracking problems in image sequences up to now.

According to Cressie and Wikle [3], a spatio-temporal marked point process can be viewed as an extension of the spatial marked point process to the temporal domain. In their view, one can think of a spatio-temporal point process either as a point process in $\mathbb{R}^{d+1}$ which they call descriptive, or as a temporal process of spatial point processes, which they call dynamic. Diggle et al. [8] gives an extensive review of the growing literature in spatio-temporal models. Nevertheless, we emphasize a change in paradigm between the spatio-temporal marked point process models found in literature and our approach.

The models, as presented by Cressie and Wikle, are used to obtain the posterior distribution of all unknowns, given the spatio-temporal information. Thus, the aim is to identify and understand the forces that drive the evolution of a certain event. This knowledge can further be used to ex-

plain the evolution of similar events. Applications include stochastic models of biological growth [13], the spread of infectious diseases [9] or the evolution of forest fires [16].
Our line of reason is exactly the opposite. The forces that drive the dynamics of the considered event are known a priori and integrated into an internal energy term. The aim in this case is to isolate and group the data that is best explained by our model from a large set of information.
We propose a new spatio-temporal marked point process model specifically adapted to the problem of multiple target tracking. We use ellipses to model the objects, *i.e.* biological particles or boats, and add an additional mark to facilitate the association between objects in different frames. We develop a new, intuitive energy and show the high detection and good association properties. We use reversible jump Markov Chain Monte Carlo (RJMCMC) [12] to find the most likely configuration of objects. We show results on three synthetic biological image sequences of 25 frames each with varying levels of noise, as well as on two difficult sequences of 14 gray-scale high resolution frames taken by an optical satellite at different angles and low temporal frequency.
This paper is organized as follows: We describe our approach to multi-target tracking in section 2. The optimization technique is presented in section 3. Section 4 shows experimental results. Finally, conclusions are drawn in section 5.

2. Multiple target tracking

2.1. Preliminaries and notation

To facilitate understanding, we first introduce the notations used throughout this paper. We model the 3D image cube as a continuous bounded set $\mathcal{K} = [0, I_{h_{max}}] \times [0, I_{w_{max}}] \times [0, T]$ and denote $x = (c_h, c_w, t)$ a point of $\mathcal{K}$, where (c_h, c_w) denote the location of the point within the image and t is the frame number. A configuration of points $\mathbf{x}$ is an unordered set of points in $\mathcal{K}$: $\mathbf{x} = \{x_1, \ldots, x_{n(\mathbf{x})}\}, x_i \in \mathcal{K}$, where $n(\mathbf{x}) = card(\mathbf{x})$ denotes the number of points in the configuration. A point process X is a collection of random configurations on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ [18]. To describe configurations of objects, random marks are added to each point. In our case of ellipses, we consider the mark space $\mathcal{M} = [a_m, a_M] \times [b_m, b_M] \times]-\frac{\pi}{2}, \frac{\pi}{2}] \times [0, L]$, where a_m, a_M and b_m, b_M are the minimum and maximum length of the semi-major and semi-minor axis respectively, $\omega \in]-\frac{\pi}{2}, \frac{\pi}{2}]$ is the orientation of the ellipse and $l \in [0, L]$ is its label. Thus, an ellipse u can be defined as $u = (c_h, c_w, t, a, b, \omega, l)$ and a marked point process of ellipses X is a point process on $\mathcal{W} = \mathcal{K} \times \mathcal{M}$. While the semi-axes a and b and the orientation ω describe the physical properties of an ellipse, the label l is used as an identifier. Objects with the same label across the image sequence form a track. We model the likelihood that an object exists at any given location in $\mathcal{K}$ and the interaction between objects in a configuration. Individual trajectories are extracted by grouping objects according to their label. Finally, we denote with $\mathcal{C}$ the set of finite configurations of ellipses.
An attractive property of point processes is the possibility of defining a point process distribution by its probability density function where the Poisson point process plays the analogue role of the Lebesgue measure on $\mathbb{R}^d$, where d is the dimension of the object space. We use the Gibbs family of processes to define the probability density as follows:

$$f_\theta(X = \mathbf{X}|\mathbf{Y}) = \frac{1}{c(\theta|\mathbf{Y})} \exp^{-U_\theta(\mathbf{X},\mathbf{Y})} \tag{1}$$

where:

- $\mathbf{X} = \{\mathbf{x_1} \cup \mathbf{x_2} \cup \cdots \cup \mathbf{x_t} \cup \cdots \cup \mathbf{x_T}\}$ is the configuration of ellipses, with $\mathbf{x_t}$ being the configuration of ellipses at time t;
- $\mathbf{Y}$ represents the 3D image cube;
- θ is the parameter vector;
- $c(\theta|\mathbf{Y}) = \int_\Omega \exp^{-U_\theta(\mathbf{X},\mathbf{Y})} \mu(d\mathbf{X})$ is the normalizing constant, with Ω being the configuration space and $\mu(\cdot)$ being the intensity measure of the reference Poisson process;
- $U_\theta(\mathbf{X}, \mathbf{Y})$ is the energy term.

The most likely configuration of objects corresponds to the global minimum of the energy:

$$X \in arg \max_{\mathbf{X} \in \Omega} f_\theta(X = \mathbf{X}|\mathbf{Y}) = arg \min_{\mathbf{X} \in \Omega} [U_\theta(\mathbf{X}, \mathbf{Y})]. \tag{2}$$

The energy function is divided in two parts: an external energy term, $U^{ext}_{\theta_{ext}}(\mathbf{X}, \mathbf{Y})$ which determines how good the configuration fits the input sequence, and an internal energy term, $U^{int}_{\theta_{int}}(\mathbf{X})$, which incorporates knowledge about the interaction between objects in a single frame and across the entire batch considered. The total energy can be written as the sum of these two terms:

$$U_\theta(\mathbf{X}, \mathbf{Y}) = U^{ext}_{\theta_{ext}}(\mathbf{X}, \mathbf{Y}) + U^{int}_{\theta_{int}}(\mathbf{X}). \tag{3}$$

The parameter vectors of the external and internal energy terms are θ_{ext} and θ_{int} respectively and $\theta = \{\theta_{ext}, \theta_{int}\}$. In the following subsections we will describe each of these two energy parts in detail.

2.2. External energy term

The external energy term is composed of the local external energies of each object u in the configuration. In order to enhance the image evidence, two local terms are computed for each object: an object evidence and a contrast distance measure.

2.2.1 Object evidence.

We search for likely locations of moving objects by frame differencing. At each pixel location, we compute the mean over time of the radiometric values and denote it p_m. Next, for each frame f, for all pixels belonging to frame f, we compute the difference between their radiometric value p_f and the mean value p_m at that location. Finally, we retain as foreground only those pixels for which this difference is higher than a predefined threshold: $\forall f \in [0,T], \forall p_f \in f : p_f \in foreground \iff |p_f - p_m| \geq threshold$. Morphological erosion and closing operations are used to enhance the filter response and smooth the boundaries of the foreground regions. We can define the class of a pixel p as $\nu(p) = \{foreground, background\}$. We compute the evidence of object u in the following way:

$$\mathcal{E}(u|\mathbf{Y}) = -\frac{1}{|u|}\sum_{p\in u}\mathbb{1}\{\nu(p) = foreground|\mathbf{Y}\} \quad (4)$$

where $|u|$ marks the cardinality of object u (e.g. the number of pixels that belong to u) and $\mathbb{1}\{\cdot\}$ marks the indicator function ($\mathbb{1}\{true\} = 1$, $\mathbb{1}\{false\} = 0$). The object evidence $\mathcal{E}(u|\mathbf{Y})$ is used to favor the detection of smaller objects.

2.2.2 Contrast distance measure.

The aim of this term is to further refine the detection and extract information such as the orientation and the size of the objects. The objects of interest (e.g. boats) appear as bright structures on a dark background. Hence, a contrast distance measure is computed between the interior of the ellipse and its border. The contrast distance measure was first introduced in [11] and is defined as:

$$d_B(u, \mathcal{F}^\rho(u)) = \frac{(\mu_u - \mu_\mathcal{F})^2}{4\sqrt{\sigma_u^2 + \sigma_\mathcal{F}^2}} - \frac{1}{2}\log\Big(\frac{2\sqrt{\sigma_u^2\sigma_\mathcal{F}^2}}{\sigma_u^2 + \sigma_\mathcal{F}^2}\Big) \quad (5)$$

where (μ_u, σ_u^2) and $(\mu_\mathcal{F}, \sigma_\mathcal{F}^2)$ represent empirical means and variances of the object u and its ρ-wide border $\mathcal{F}^\rho(u)$. A threshold $d_0(\mathbf{Y})$ is manually determined based on the image. High threshold values are used when the objects are easily distinguishable from the background. Lower threshold values are used otherwise.
A quality function $\mathcal{Q} : \mathbb{R}^+ \to [-1, 1]$ is used to compensate for errors close to the threshold value:

$$\mathcal{Q}(x) = \begin{cases} 1 - x^{1/3} & \text{if } x < 1 \\ exp(-\frac{x-1}{3}) - 1 & \text{if } x \geq 1 \end{cases} \quad (6)$$

The quality function attributes a negative value to well placed ellipses (e.g. objects u for which $d_B(u, \mathcal{F}^\rho(u))$ is higher than the threshold $d_0(\mathbf{Y})$) and a positive value otherwise.

The two terms computed in eq. 4 and eq. 6 are further combined into a local external energy for an object u:

$$U_{local}^{ext}(u|\mathbf{Y}) = \gamma_{ev}\mathcal{E}(u|\mathbf{Y}) + \gamma_{cnt}\mathcal{Q}\Big(\frac{d_B(u, \mathcal{F}^\rho(u))}{d_0(\mathbf{Y})}\Big). \quad (7)$$

Finally, the external energy term for the configuration $\mathbf{X}$ is:

$$U_{\theta_{ext}}^{ext}(\mathbf{X}, \mathbf{Y}) = \sum_{u\in\mathbf{X}} U_{local}^{ext}(u|\mathbf{Y}). \quad (8)$$

The parameter vector $\theta_{ext} = \{\gamma_{ev}, \gamma_{cnt}\}$ of the external term consists of the weight γ_{ev} of the evidence term and the weight γ_{cnt} of the quality of the contrast distance.

2.3. Internal energy term

The internal energy term consists of a set of constraints meant for a correct detection of objects and to facilitate tracking. These constraints are inspired by the physical constraints objects obey in real life.

2.3.1 The dynamic model.

A defining property of tracking (as opposed to individual detections per frame) is that in most cases object trajectories are smooth. This allows to favor configurations where objects exhibit a motion described by a dynamic model. This motion model, denoted by dyn, depends on the application. Nevertheless, we can create an energy term that encourages objects to follow a given motion model s.t. for an object u that exists at time t it can be written as:

$$U_{dyn}^{int}(u) = \begin{cases} dyn_0 - dyn & \text{if } \exists\ dyn \text{ s.t. } dyn \leq dyn_0 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where dyn_0 is a threshold that describes how much objects can deviate from the motion model and still be awarded.
The energy term that awards configurations which follow the dynamic model is the sum over all objects in the configuration:

$$U_{dyn}^{int}(\mathbf{X}) = \gamma_{dyn}\sum_{u\in\mathbf{X}} U_{dyn}^{int}(u). \quad (10)$$

This term favors the creation of objects where the data evidence is reduced but the dynamic model motivates the existence of an object.

2.3.2 Label persistence.

In order to distinguish between distinct trajectories we have introduced a label in the mark of each object. A label can be viewed as a trajectory identifier. Different labels mean different trajectories. Thus, the number of labels has to be kept closely related to the number of trajectories in the data set. Ideally, the large number of objects u scattered across the

image sequence should be assigned to a rather small number of labels. In this regard, we construct the set of labels present in a configuration $\mathbf{X}$ by $labels(\mathbf{X}) = \bigcup_{u \in \mathbf{X}} l(u)$, where $l(u)$ is the label of object u. We favor configurations where the number of distinct labels is small:

$$U^{int}_{label}(\mathbf{X}) = -\gamma_{label}\Big(\frac{1}{|labels(\mathbf{X})|}\Big) \tag{11}$$

where $|labels(\mathbf{X})|$ represents the cardinality of the set.
The labels are assigned based on the motion model. Given object u centered at location $pos(u) = (c_h(u), c_w(u))$, we search for the objects in the adjacent frames that satisfy the motion model. We compute the distance between u and these objects and compare it to a threshold. If the distance is smaller than the threshold, we set the label of object u to the label of the object in the previous frame. Otherwise, a new random label from $[0, L] \setminus labels(\mathbf{x_t})$ is assigned to u. Configurations $\mathbf{X}$ that contain two or more objects with the same label at any time instance t are not permitted, meaning that an infinite energy is assigned to such configurations.

2.3.3 Mutual exclusion.

Handling object collision or overlapping at a given frame is a crucial aspect when detecting and tracking objects. In our model, we attribute an infinite penalty to any configuration that contains objects that overlap more than a given extent s. Thus, the probability of selecting such a configuration is zero. We denote by

$$A(u, v) = \frac{Area(u \cap v)}{\min(Area(u), Area(v))} \tag{12}$$

the area of intersection between the objects u and v. The energy term describing the penalty for overlapping is

$$U^{int}_{overlap}(\mathbf{X}) = \begin{cases} \sum_{u,v \in \mathbf{X}, u \neq v} A(u, v) \\ \quad \text{if } \forall t \in [0, T] \forall u, v \in \mathbf{x}_t : A(u, v) \leq s \\ +\infty \quad \text{otherwise.} \end{cases} \tag{13}$$

The reason for which we choose to impose this hard constraint is the fact that our main data set is composed of remotely sensed images. Since our interest lies in detecting and tracking real-life objects, we can fairly conclude that two distinct objects cannot simultaneously occupy the same image region at any time instance.

2.4. Total energy term

The total energy term can be written as a sum of all the energy terms defined in section 2.2 and section 2.3:

$$\begin{aligned} U_\theta(\mathbf{X}, \mathbf{Y}) \quad = \quad & \gamma_{dyn} U^{int}_{dyn}(\mathbf{X}) + \gamma_{label} U^{int}_{label}(\mathbf{X}) + \\ & \gamma_o U^{int}_{overlap}(\mathbf{X}) + \gamma_{ev} \mathcal{E}(u|\mathbf{Y}) + \\ & \gamma_{cnt} \sum_{u \in \mathbf{X}} \Big(\mathcal{Q}\Big(\tfrac{d_B(u, \mathcal{F}^\rho(u))}{d_0(\mathbf{Y})}\Big)\Big). \end{aligned} \tag{14}$$

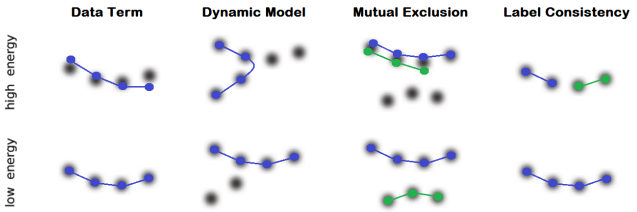

Figure 1. The effects of different components of the energy terms. The upper row shows a configuration with a higher energy value for each individual term. The bottom row shows a configuration with a lower energy value for each individual term. The dark spots denote target locations at different time frames. Different colors on the targets represent different labels assigned to each.

An intuition of how each energy term influences the output result is presented in Figure 1.

2.5. Parameters

The weights of the energy terms present in eq. 14 do not have an intuitive meaning and thus are harder to set by hand. Therefore, an automatic method to determine these parameters is necessary. In our experiments we have used a linear programming approach to estimate these weights. Given a configuration $\mathbf{X}$, the log posterior density is a linear combination of the parameters which can be considered to be independent from each other. Although we do not have access to the precise value of the log posterior density due to the normalizing constant, we can however compute the ratio, r, between two log posterior densities. If we know that one configuration is better than another, we can establish a set of constraints $r \leq 1$ (or $r \geq 1$). These constraints can be transformed into a set of linear inequalities of the parameters. Once this set of inequalities is large enough, we can apply linear programming to obtain a feasible solution for the parameters. The interested reader can refer to [21] for further details. Alternative parameter estimation techniques are discussed in [7].

3. Optimization

The energy described in eq. 14 is clearly not convex. It is easy to construct examples that have two virtually equal minima, separated by a wall of high energy values. The dependence caused by the high-order physical constraints is the main reason that drives the energy to be non-convex.
The target distribution is the posterior distribution of $\mathbf{X}$, *i.e.* $\pi(\mathbf{X}) = f(\mathbf{X}|\mathbf{Y})$, defined on a union of subspaces of different dimensions. The most widely known optimization method for non-convex energy functions and an unknown number of objects is the reversible jump Markov Chain Monte Carlo (RJMCMC) sampler developed by Green [12]. RJMCMC uses a mixture of perturbation kernels $Q(\cdot, \cdot) = \sum_m p_m Q_m(\cdot, \cdot)$, $\sum_m p_m = 1$ and $\int Q_m(\mathbf{X}, \mathbf{X}')\mu(d\mathbf{X}') =$

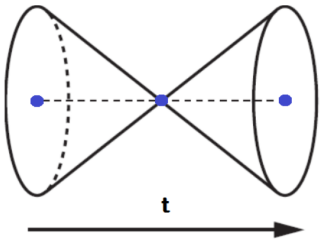

Figure 2. The space-time volume an object can physically influence from its current position at time t can be depicted as a double-cone extending both in frame $t-1$ and $t+1$.

1, to create tunnels through the walls of high energy.
We use simulated annealing to find a minimizer of the energy function. The density function in eq. 1 can is rewritten as:

$$f_{\theta,i}(X=\mathbf{X}|\mathbf{Y}) = \frac{1}{c_{Temp_i}(\theta|\mathbf{Y})} \exp^{-\frac{U_\theta(\mathbf{X},\mathbf{Y})}{Temp_i}} \qquad (15)$$

where $Temp_i$ is a temperature parameter that tends to zero when i tends to infinity. If $Temp_i$ decreases in logarithmic rate, then X_i tends to a global optimizer of $f_{\theta,i}$. In practice however, a logarithmic law is not computationally feasible and hence, a geometric law is used instead. Therefore, a proper design of the perturbation kernels is needed to ensure a good exploration of the state space.

The efficiency of this iterative algorithm depends on the variety of the perturbation kernels. We have used the following perturbation kernels in our experiments:

- *Birth and death according to a birth map*: Two types of maps are created in a pre-processing step:

  1. *Birth maps*: since the objects are supposed to have higher radiometric values than the background, we use a simple threshold technique to identify probable locations of objects in each frame and attribute higher probabilities to these locations for the birth proposition kernel;
  2. *Water mask*: in the case of boat tracking we first detect the water area and limit the search to such areas. Crăciun et al. [2] present a simple and effective method to extract the water area based on three features that characterize the water area: low radiometric values, small variance across the area and a relative large size. A single water mask is computed for the whole image sequence;

  The birth and death according to a birth map kernel first chooses with probability p_b and $p_d = 1-p_b$ whether an object u should be added to (birth) or deleted from (death) the configuration. If a birth is chosen, the kernel generates a new object u according to the birth map and proposes $\mathbf{X}' = \mathbf{X} \cup u$. If a death is chosen, the kernel selects one object u in $\mathbf{X}$ according to the birth map and proposes $\mathbf{X}' = \mathbf{X} \setminus u$;

- *Birth and death in a neighborhood*: this kernel is used to propose the addition or removal of an interacting pair of objects. To define the neighborhood of an object we introduce the notion of *event cones*. This notion was previously introduced by Leibe et al. [14] to search for plausible trajectories in the space-time volume by linking up event cones. Following the idea of Leibe et al. we define the event cone of an object $u = (c_h, c_w, t, a, b, \omega, l)$ to be the space-time volume it can physically influence from its current position as depicted in Figure 2;

- *Non-jumping transformations*: non-jumping transformations are transformations that randomly select an object u in the current configuration and then propose to replace it by a perturbed version of the object v: $\mathbf{X}' = (\mathbf{X} \setminus u) \cup v$. Translation, rotation and scale are examples of such transformations.

A mapping $R_m(\cdot,\cdot) : \mathcal{C} \times \mathcal{C} \to (0,\infty)$, called the Green ratio, is associated to each of these perturbation kernels. At iteration i, the proposition $X_i = \mathbf{X}'$ is accepted with probability $\alpha_m = \min(1, R_m(\mathbf{X}, \mathbf{X}'))$. Otherwise $X_i = \mathbf{X}$. Although embedding the RJMCMC sampler into a simulated annealing scheme yields better results, in practice it is computationally very expensive.

4. Results and discussions

We first test our approach on three synthetic biological image sequences. The image sequences have been generated with Icy [6], an online available toolbox for biological image sequences developed by the Pasteur Institute in Paris. The sequences consist of 25 images, each 256×256 pixels, with approximately 20 objects per frame. The three sequences exhibit different levels of Gaussian noise. The tracking results are displayed in Figure 3. The motion model of the objects is considered to be a Brownian motion.
We compare our approach with the built-in particle tracker that comes as a plug-in to the Icy software. The results are shown in Table 1. Nevertheless, the output of the MHT tracker should not be considered as ground truth. The similarities between the paired detections and tracks are computed based on a maximum distance of 5 pixels between the two methods. This means that if the same object is detected using MHT at position $c1$ and with the proposed approach at position $c2$ and if $|c1 - c2| \geq 5$ then the two detections are not matched.
The computation time for each sequence is around 1 minute

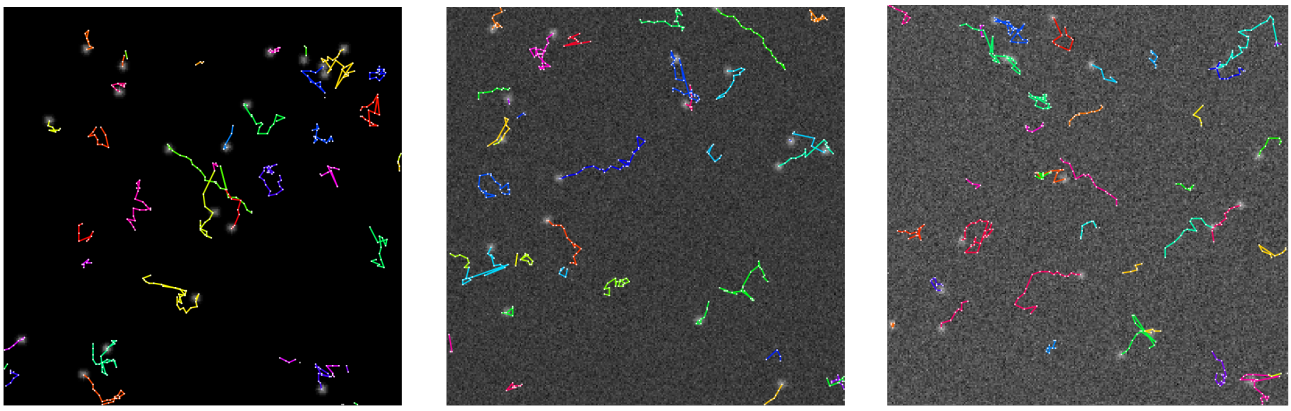

Figure 3. Detection and tracking results on synthetic biological image sequences created using Icy software [6]. Left: Tracking results on the first image sequence (no noise). Middle: Tracking results on the first image sequence (Gaussian noise, $\mu = 25$, $\sigma = 2.5$). Right: Tracking results on the first image sequence (Gaussian noise, $\mu = 50$, $\sigma = 5.0$).

Data set	No. reference tracks (MHT)	No. candidate tracks (Proposed algorithm)	Similarity between tracks	Similarity between detections
Seq. 1	45	53	0.3803	0.4230
Seq. 2	49	38	0.4263	0.3285
Seq. 3	49	40	0.3485	0.2674

Table 1. Comparison between the results obtained using the built-in MHT tracker within Icy [6] and the proposed method for the three synthetic image sequences. Note however, that the output of the MHT tracker should not be taken as ground truth information.

on a $2.30GHz$ Intel Xeon processor. Although this performance is reduced compared to other state-of-the-art methods, we believe that the performance of our algorithm can be greatly increased using a parallel computing scheme. The implementations developed by Verdie et al. [19] and Crăciun et al. [2] have brought significant speed-up in the case of object detection in a single image. A similar scheme can be envisioned for the dynamic case of object tracking over a sequence of images.

Our main objective however, is to apply the proposed algorithm on real data. The acquisition rate of satellite images has experienced a significant increase in the last years. Therefore, object tracking using high resolution satellite images can be regarded as a new application in remote sensing, complementary to object detection and land-cover classification. Therefore, we test our approach on two challenging image sequences of boats. Each sequence consists of 14 frames taken at a low temporal frequency. Targets exhibit strong variations in appearance due to the changing angle at which the images were taken. We compare our results to two classical trackers: Kalman filter [20] and Histogram-Based Tracker [5]. We consider a constant velocity motion model in this case.

Metrics. Conducting an objective comparison between different tracking algorithms is a challenging task for various reasons. First, the importance of individual tracking failures is application dependent. Second, classifying tracker outputs as correct or incorrect may as well be very ambiguous and usually requires additional parameters (e.g. thresholds) to assess the correctness and precision of the trackers.

To evaluate the multi-object tracking accuracy, we compute three types of errors: false positives (FP), false negatives (FN) and identity switches (ID). The three types of errors are weighted equally. We also state the number of true positives (TP) and we provide the total number of moving objects (TO). The total number of moving objects (TO) is the sum over all frames of the objects that change their position in two consecutive frames. Additionally, mostly tracked (MT) and mostly lost (ML) scores are computed on the entire number of distinct trajectories (TT) to measure how many ground truth trajectories are tracked successfully (tracked for at least 80 percent) or lost (tracked for less than 20 percent). Finally, we state the precision (TP / (TP + FP)) and recall (TP / (TP + FN)) of each algorithm.

Quantitative evaluation. Table 2 shows the quantitative results for both image sequences individually. We show the results of three trackers: our full model including dynamic birth maps and the water mask used for optimiza-

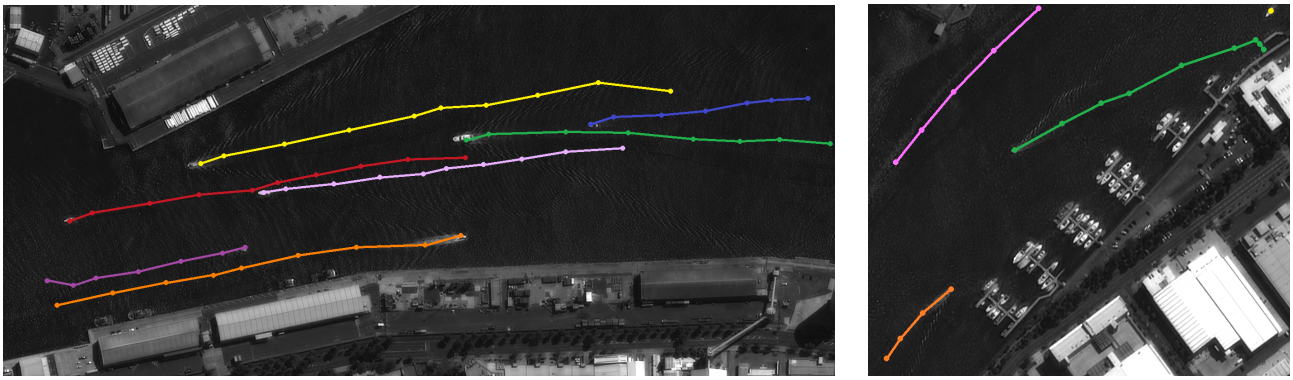

Figure 4. Detection and tracking results on two sequences of satellite images taken at different angles. Left: Tracking results on the first image sequence up to frame 10. Right: Tracking results up to frame 13 of the second image sequence.

Data set	Method	FP	FN	TP	TO	ID	MT	ML	TT	Precision	Recall
Seq. 1	ST-MPP + BM	1	6	85	91	0	7	1	8	**0.988**	0.934
	KF	3	34	57	91	0	4	2	8	0.950	0.626
	HBT	5	14	77	91	3	6	2	8	0.939	0.846
	MPP	7	5	84	91	–	–	–	–	0.923	**0.944**
Seq. 2	ST-MPP + BM	1	1	24	25	0	4	0	4	**0.962**	**0.962**
	KF	2	4	21	25	0	2	1	4	0.913	0.84
	HBT	2	3	22	25	1	2	0	4	0.916	0.88
	MPP	3	1	24	25	–	–	–	–	0.889	0.961

Table 2. Quantitative results for the two sequences of satellite images.

tion, denoted (ST-MPP + BM), the Kalman filter (KF) and the histogram-based tracker (HBT). Figure 4 shows the results of our method on the two image sequences considered. For comparison, we list in Table 2 also the detection results of the spatial marked point processes model developed by Crăciun et al. [4] and denoted (MPP) which we applied independently on each frame to extract boats.

The first image sequence has 14 frames of 1840×820 pixels each and contains a constant number of 8 moving objects throughout the entire sequence. The use of the object evidence term leads to better estimates of the locations of the objects, while the contrast distance measure is used to obtain accurate values for the size and orientation of the objects. Moreover, the labels are preserved throughout the image sequence. The computation lasted in average 20 minutes on a $2.30GHz$ Intel Xeon processor. The second image sequence has 14 frames of 830×730 pixels each. The evidence term plays a decisive role in distinguishing dynamic objects from static ones. Our model yields a higher tracking performance compared to the classical trackers, due to the better detection results. The labels of the objects are generally preserved throughout the sequence. The average computation lasted in average 7 minutes on the same $2.30GHz$ Intel Xeon processor.

Our method (ST-MPP + BM) outperforms both the Kalman filter (KF) and the histogram-based tracker (HBT). The lower performance of the Kalman filter is described by the lower performance of the detector used and the time it needs for initialization. The performance of the histogram-based tracker is highly influenced by the change in illumination due to the different angles of acquisition of the images. The appearance of the objects changes throughout the sequence and thus, the precision of the tracker is affected. In terms of appearance, our tracker however only relies on the contrast between the objects and their border. Therefore, its performance is not affected by appearance changes.

5. Conclusions and future work

In this paper we have presented a novel spatio-temporal marked point process of ellipses to detect and track moving boats in high resolution remotely sensed images sequences. First, we have emphasized the ideological difference between current spatio-temporal marked point processes designed for statistical understanding of natural events and our model designed for detecting and tracking moving objects in image sequences. Then, we have defined a new and intuitive energy to detect the objects across the sequence and group them into trajectories. We have used an adapted

version the widely known RJMCMC sampler for optimization. We have computed birth maps for each frame and proposed the use of the birth and death in the neighborhood permutation kernel to speed up the optimization process. We have then tested our algorithm on three synthetic biological sequences with varying levels of noise. Finally, we have shown promising results on two remotely sensed high resolution optical images sequences.

Future work can be envisioned along three complementary directions: first, an in-depth analysis of the robustness of the model with respect to noise and outliers has to be performed. Moreover, the model could be further extended to include the detection of static objects. Second, a parallel implementation of the RJMCMC sampler for spatio-temporal marked point process models needs to be devised. Such an implementation would significantly decrease the computational time of the sampler. A data-parallel implementation based on the conditional independence of targets that are far apart within a frame can be envisioned. Such an approach would minimize the communication cost between clusters. Finally, the model could be applied in other fields, such as videomicroscopy.

Acknowledgments

The authors would like to thank Airbus D&S (http://airbusdefenceandspace.com) for providing the satellite data and for the partial funding of this research. We would like to thank Pasteur Institute (www.pasteur.fr) for the free access to the Icy software for biological image processing (http://icy.bioimageanalysis.org).

References

[1] Y. Bar-Shalom and T. Fortmann. *Tracking and Data Association*. Academic Press, San Diego, 1988.

[2] P. Craciun and J. Zerubia. Towards efficient simulation of marked point process models for boat extraction from high resolution optical remotely sensed images. *Proc. IGARSS*, 2014.

[3] N. Cressie and C. Wikle. *Statistics for spatio-temporal data*. Wiley series in probabilities and statistics, 2011.

[4] P. Crăciun and J. Zerubia. Unsupervised marked point process model for boat extraction in harbors from high resolution optical remotely sensed image. *Proc. ICIP*, pages 4122–4125, 2013.

[5] N. Dalai and B. Triggs. Histograms of oriented gradients for human detection. *Proc. CVPR*, 1:886–893, 2005.

[6] F. de Chaumont et al. Icy: an open bioimage informatics platform for extended reproducible research. *Nature methods*, 9:690–696, 2012.

[7] X. Descombes, F. Chatelain, F. Lafarge, C. Lantuejoul, C. Mallet, M. Minlos, M. Schmitt, M. Sigelle, R. Stoica, and E. Zhizhina. *Stochastic Geometry for Image Analysis*. John Wiley and Sons, 2011.

[8] P. Diggle. Spatio-temporal point processes: methods and applications. *Statistical methods for spatio-temporal systems*, pages 1–45, 2007.

[9] P. Diggle, B. Rowlingson, and T.-L. Su. Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, 16:423–434, 2005.

[10] W. Ge and R. T. Collins. Marked point processes for crowd counting. *Proc. CVPR*, pages 2913–2920, 2009.

[11] F. Goudail, P. Réfrégier, and G. Delyon. Bhattacharyya distance as a contrast parameter for statistical processing of noisy optical images. *Journal of Optical Science of America A*, 21(7):1231–1240, 2004.

[12] P. Green. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[13] E. Jensen, K. Jónsdóttir, J. Schmiegel, and O. Barndoff-Nielsen. Spatio-temporal modeling - with a view to biological growth. *Statistical methods for spatio-temporal systems*, pages 47–75, 2007.

[14] B. Leibe, K. Schindler, N. Cornelis, and L. van Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE TPAMI*, 30(10):1683–1698, 2008.

[15] S. Oh, S. Russell, and S. Sastry. Markov Chain Monte Carlo Data Association for general multiple-target tracking problems. *Proc. CDC*, 1:735–742, 2004.

[16] R. Peng, F. Schoenberg, and J. Woods. A space-tile conditional intensity model for evaluating a wildfire hazard index. *J. Amer. Statist. Assoc.*, 100:26–35, 2005.

[17] K. Smith, A. Carleton, and V. Lepetit. General constraints for batch multiple-target tracking applied to large-scale videomicroscopy. *Proc. CVPR*, pages 1–8, 2008.

[18] M. van Lieshout. *Markov Point Processes and Their Applications*. Imperial College Press, 2000.

[19] Y. Verdié and F. Lafarge. Efficient Monte Carlo sampler for detecting parametric objects in large scenes. *Proc. ECCV*, 7574:539–552, 2012.

[20] G. Welch and G. Bishop. An introduction to the Kalman filter. *Proc. SIGGRAPH*, pages 19–24, 2001.

[21] Q. Yu and G. Medioni. Multiple-target tracking by spatio-temporal Monte Carlo Markov chain data association. *IEEE TPAMI*, 31(12):2196–2210, 2009.