Hybrid Stochastic / Deterministic Optimization for Tracking Sports Players and Pedestrians*

Robert T. Collins[1] and Peter Carr[2]

[1] The Pennsylvania State University, USA
[2] Disney Research Pittsburgh, USA

Abstract. Although 'tracking-by-detection' is a popular approach when reliable object detectors are available, missed detections remain a difficult hurdle to overcome. We present a hybrid stochastic/deterministic optimization scheme that uses RJMCMC to perform stochastic search over the space of detection configurations, interleaved with deterministic computation of the optimal multi-frame data association for each proposed detection hypothesis. Since object trajectories do not need to be estimated directly by the sampler, our approach is more efficient than traditional MCMCDA techniques. Moreover, our holistic formulation is able to generate longer, more reliable trajectories than baseline tracking-by-detection approaches in challenging multi-target scenarios.

1 Introduction

Multi-target tracking of pedestrians and sports players is difficult due to the presence of many similar-looking objects interacting in close proximity. For this reason there has been recent interest in sliding temporal window methods that recover tracking solutions by considering a batch of frames at a time. The motivation is that people who are occluded or otherwise difficult to disambiguate in a few frames will be easier to find in others, and that propagating temporal consistency constraints both backwards and forwards in time leads to better solutions than purely causal processing.

It is also advantageous to solve for detections and data association jointly, rather than computing detections first and then linking them into trajectories. Despite the obvious benefits, this holistic approach has received considerably less attention because the complexity of the search space of data association increases exponentially with the number of candidate detections in each frame, and therefore committing to a small set of high-quality discrete detections makes the later association problem more manageable. However, not being able to reconsider detection decisions puts a large burden on the data association algorithm to handle deficiencies such as missed detections and false positives.

* Electronic supplementary material -Supplementary material is available in the online version of this chapter at http://dx.doi.org/10.1007/978-3-319-10605-2_20. Videos can also be accessed at http://www.springerimages.com/videos/978-3-319-10604-5.

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 298–313, 2014.

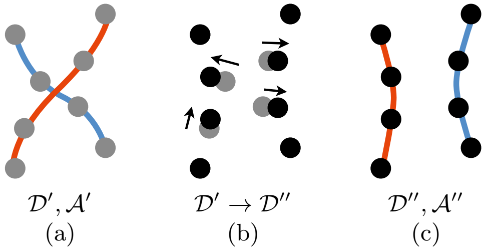
$\mathcal{D}', \mathcal{A}'$
(a)
$\mathcal{D}' \to \mathcal{D}''$
(b)
$\mathcal{D}'', \mathcal{A}''$
(c)

Fig. 1. Stochastic Detection/Deterministic Tracking. (a) An initial set of detections $\mathcal{D}'$ has a corresponding optimal data association solution $\mathcal{A}'$, shown here as red and blue trajectories. However, due to detection noise, we may have mistakenly swapped the identities of the two targets. (b) If we stochastically perturb the set of detections to generate a new hypothesis $\mathcal{D}''$, it may lead (c) to a better data association solution $\mathcal{A}''$. Conceptually, we are decomposing the joint optimization of $(\mathcal{D}, \mathcal{A})$ into a stochastic proposal of multi-frame detections $\mathcal{D}$ and a deterministic solution for $\mathcal{A}|\mathcal{D}$ (similar to 'line search') given each such proposal.

In this paper we present a Bayesian approach for simultaneous optimization over the space of detections and data associations (Figure 1). We develop a hybrid optimization algorithm that uses Reversible Jump Markov Chain Monte Carlo (RJMCMC) sampling over the space of detections to "drive" the estimation process, while leveraging recent results on deterministic polynomial-time algorithms for computing the globally optimal data association for a set of hypothesized multi-frame detections. Experimental results show that the method performs well, even on sports sequences where players perform rapid maneuvers in close proximity to each other.

Contributions

1) Our main contribution is a hybrid MCMC algorithm that uses deterministic solutions for blocks of variables to accelerate its stochastic mode-seeking behavior. Incorporating deterministic solutions within MCMC is nearly universally avoided [18] because it breaks detailed balance and threatens the integrity of the sampler. However, we note that using MCMC to guide discovery of the joint mode of a posterior distribution does not require faithful generation of samples representative of the whole distribution, and show that factoring a joint distribution into detection and association variables leads to a natural framework where MCMC sampling over detections is interleaved with a deterministic solution for the optimal set of associations. We show that our method yields a correct sampler with respect to a max-marginal distribution over detections, and that seeking the mode of this max-marginal allows efficient search for the joint mode of the original posterior with respect to both detections and associations.

2) Unlike the majority of tracking-by-detection methods for multi-frame, multi-target tracking, our approach iteratively revises (including adding and removing) detections over the sequence of frames. This leads to better results than fixing a set of detections once and for all prior to performing data association. Furthermore, interleaving data association with the search for detections has a

regularizing effect that encourages consistency of the number of detections across frames and of their locations with respect to a smooth path, without having to enforce those qualities through trajectory/motion smoothness priors.

2 Related Work

Tracking-by-detection [16,28,5,9,23,2,27] is a popular technique for multi-object tracking. A sequence of frames is preprocessed with an object detector to generate a finite set of object locations in space/time, and data association is then used to link detections across discrete time intervals, which effectively estimates a sampled trajectory for each object (to which a smoothed approximation can be estimated [2]). When the association objective function is limited to addition or multiplication of pairwise costs, the global optimum can be computed in polynomial time [5,23]. Most methods can easily handle false detections, but missed detections are more difficult since links must be hypothesized to span multi-frame gaps.

Breitenstein *et al.* [8] use detector confidence maps to hypothesize new locations when detections are missing. Our approach is similar to [8] in that we also hypothesize detections, however we hypothesize **all detections** for the entire multi-frame sequence, and not just detections which may have been missed by an object detector. Other approaches similar to ours, in that they attempt to simultaneously estimate both detections and trajectories, include: [19], where combined detection and trajectory estimation becomes an NP-hard quadratic boolean optimization problem, solved heuristically; the non-convex continuous energy minimization approach of [21], which contains transdimensional jump moves similar to RJMCMC, although applied in a deterministic way that can only decrease the energy; and [26], who propose a coupled detection and tracking approach where a sparsity-constrained detection solution is interleaved with min-cost flow data association in a Lagrangian optimization loop.

Markov Chain Monte Carlo (MCMC) sampling methods offer a general approach for exploring large problem spaces under expressive objective functions, and have been applied to problems of multi-target detection [29,13] and data association [20,7]. Previous MCMC Data Association (MCMCDA) approaches [14,22,4,20,7,17] have explicitly estimated the associations between detections. However, not only is the space of unknowns to be explored much larger when assignment links are included in the MCMC search, one has to design specialized moves that propose coordinated changes to multiple assignment variables to satisfy the one-to-one matching constraints necessary to maintain a feasible solution. A key difference of our work is that we do not explicitly estimate the data association variables using stochastic search. Instead, we address data association as a closed form solution contingent on the current hypothesized set of detections. As a result, we only need to consider relatively simple and well-understood sampler moves related to detections (e.g. birth, death and diffusion).

3 Approach

In multi-target tracking, the variables to be solved for are the number and location of objects (detections) in each frame of the sequence, and the inter-frame correspondences (associations) of those detections over time to form a set of trajectories.

3.1 Bayesian Formulation

We adopt a Bayesian approach where detections $\mathcal{D}$ and associations $\mathcal{A}$ are random variables, likelihood functions measure how well a hypothetical set of detections and associations explain the observed image sequence $\mathcal{Z}$, and priors encourage properties expected in "good" solutions. The goal is to maximize the joint posterior distribution over $\mathcal{D}$ and $\mathcal{A}$ given observations $\mathcal{Z}$:

$$A^*, D^* = \underset{\mathcal{A},\mathcal{D}}{\text{argmax}}\, P(\mathcal{A},\mathcal{D}\,|\mathcal{Z}) \tag{1}$$

$$= \underset{\mathcal{A},\mathcal{D}}{\text{argmax}}\, P(\mathcal{A}\,|\mathcal{D},\mathcal{Z})P(\mathcal{D}\,|\mathcal{Z}) \tag{2}$$

where the second line follows by the definition of conditional probability.

Without loss of generality, we split the argmax and rewrite Eq. (2) as :

$$D^* = \underset{\mathcal{D}}{\text{argmax}}\left[\left(\underset{\mathcal{A}}{\text{max}}\, P(\mathcal{A}\,|\mathcal{D},\mathcal{Z})\right)P(\mathcal{D}\,|\mathcal{Z})\right] \tag{3}$$

$$A^* = \underset{\mathcal{A}}{\text{argmax}}\, P(\mathcal{A}\,|D^*,\mathcal{Z}) \ . \tag{4}$$

In practice the argmax A^* is found while computing the max over $\mathcal{A}$ in the inner parentheses of Eq. (3). This is equivalent to the joint maximization in Eq. (1) because both $P(\mathcal{A}\,|\mathcal{D},\mathcal{Z})$ and $P(\mathcal{D}\,|\mathcal{Z})$ are non-negative by construction. Intuitively, this factors the joint estimation problem into detections, $P(\mathcal{D}|\mathcal{Z})$, and data associations, $P(\mathcal{A}|\mathcal{D},\mathcal{Z})$. See Figure 2.

Previous tracking-by-detection approaches compute the following approximate solution to Eqs. (3–4):

$$D^* = \underset{\mathcal{D}}{\text{argmax}}\, P(\mathcal{D}\,|\mathcal{Z}) \tag{5}$$

$$A^* = \underset{\mathcal{A}}{\text{argmax}}\, P(\mathcal{A}\,|D^*,\mathcal{Z}) \ . \tag{6}$$

This is suboptimal even if the correct marginal $P(\mathcal{D}|\mathcal{Z}) = \int_{\mathcal{A}} P(\mathcal{A},\mathcal{D}\,|\mathcal{Z})$ is used, because the mode of a marginal distribution does not necessarily correspond to the projection of the mode of the joint distribution. Furthermore, generating a fixed set of detections prior to determining associations makes it difficult if not impossible to recover when detections are missed due to occlusion or low detector confidence. It is better to allow association-based information such as

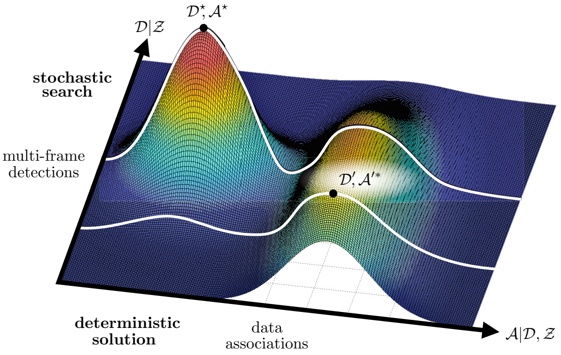
$\mathcal{D}^*, \mathcal{A}^*$
$\mathcal{D}|\mathcal{Z}$
stochastic search
multi-frame detections
$\mathcal{D}', \mathcal{A}'^*$
deterministic solution
data associations
$\mathcal{A}|\mathcal{D}, \mathcal{Z}$

Fig. 2. Hybrid Stochastic/Deterministic Optimization. The goal is to determine the optimal set of detections $\mathcal{D}^*$ and associations $\mathcal{A}^*$ for observations $\mathcal{Z}$. We factor the joint optimization into stochastic search over detections $\mathcal{D}|\mathcal{Z}$ interleaved with deterministic solutions for associations $\mathcal{A}|\mathcal{D}, \mathcal{Z}$. Each hypothesized set of detections $\mathcal{D}'$ results in a reduced 'line search' for the corresponding best set of associations $\mathcal{A}'$ (which has a deterministic solution for energy functions of pairwise potentials).

high-confidence partial trajectories to guide estimation of hard-to-see detections during ambiguous portions of the sequence.

On the other hand, Eqs. (3–4) suggests that an algorithm for estimating the joint mode can be organized as a search over multi-frame configurations of detections while using a subroutine to solve the global data association problem for each hypothesized set of detections. That is the strategy taken in this paper: we present a hybrid optimization approach that uses RJMCMC to perform stochastic search over the space of detection configurations, interleaved with deterministic computation of the optimal multi-frame assignment for each proposed detection hypothesis.

The clearest way to think about our approach is to consider deterministic computation of assignments to be a closed-form function $A(\mathcal{D})$, and that we are performing stochastic optimization over a distribution $\Psi(\mathcal{D}) \propto A(\mathcal{D})P(\mathcal{D}|\mathcal{Z})$ that is a function only of detections.[1] It is not hard to recognize that $\Psi(\mathcal{D})$ is the **max-marginal** of $P(\mathcal{A}, \mathcal{D}\,|\mathcal{Z})$ computed by max'ing over $\mathcal{A}$ for each value of $\mathcal{D}$. One insight is that sampling from the max-marginal $\Psi(\mathcal{D})$ is sufficient to guide the search for (A^*, D^*), since $\Psi(\mathcal{D})$ has the same mode D^* as $P(\mathcal{A}, \mathcal{D}\,|\mathcal{Z})$, and, once D^* is found, it can be plugged into $A(\mathcal{D})$ to find A^*.

Unfortunately, $\Psi(\mathcal{D})$ is hard to sample from directly due to the implicit coupling between associations and detections. However, an MCMC sampler may propose samples from a simpler proposal distribution and rely on computation of the acceptance ratio to make sure accepted samples are distributed according to the desired target distribution. In this work we design an MCMC sampler that uses simple local updates to current detection configuration D_c to propose a new configuration D', for which the optimal data association $A' = A(D')$ is computed

[1] Note we overload $A(D)$ to refer to both the argmax as well as the value at the max.

deterministically, followed by using $\Psi(D')$ and $\Psi(D_c)$ to compute the likelihoods in the acceptance ratio that ensure $\Psi(\mathcal{D})$ is the correct target distribution of the sampler. The components of the sampler are presented below.

3.2 Observation Data

Our method uses a subsampled temporal window of N frames $I = \{I_1, I_2, \ldots, I_N\}$. Input RGB images are converted to YCbCr so that luminance information can be treated differently from chrominance. In addition to the raw pixel data, a set of binary foreground masks $F = \{F_1, F_2, \ldots, F_N\}$ is generated by background subtraction and thresholding in YCbCr color space, followed by denoising using morphological opening and dilation operators, and optional suppression of foreground data outside of a given region of interest.

To facilitate reasoning about locations of people in the ground plane, each foreground mask F_k is mapped to a monocular *occupancy proposal map* M_k such that $M_k(x, y)$ indicates the probability of ground location (x, y) being occupied. This is performed by a process similar to [10] where F_k is backprojected using camera calibration information onto 3D volume elements of the scene, followed by marginalizing over the height dimension. Together, the N triplets of color images, binary foreground masks and occupancy maps comprise the observation data $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_N\}$, with $Z_k = (I_k, F_k, M_k)$. See Figure 3 top row.

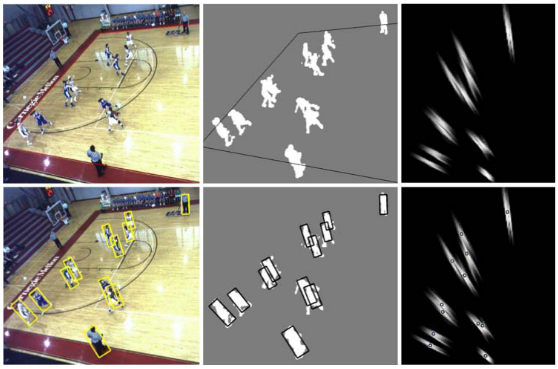

Fig. 3. Top row, left to right: color image I_k; foreground mask F_k (also showing region of interest); ground plane proposal map M_k. Bottom row: single frame detection results overlaid on each form of observation data.

3.3 Detections

A posterior distribution $P(\mathcal{D}|\mathcal{Z})$ over multi-frame detections is derived by Bayes rule and assumption of independence over frames

$$P(\mathcal{D}|\mathcal{Z}) \propto P(\mathcal{D})P(\mathcal{Z}|\mathcal{D}) = \prod_{k=1}^{N} P(D_k) \prod_{k=1}^{N} P(Z_k|D_k) \ . \tag{7}$$

Although independence is assumed, we note that later combination with the conditional posterior over associations, $P(\mathcal{A}|\mathcal{D},\mathcal{Z})$, will have a regularizing effect on the set of detections found across frames.

Each person is modeled as a 3D cylinder C with ground plane location (x, y), height h and radius r. Holding r constant, the unknowns for each detected cylinder $d_i \in C$ are (x_i, y_i, h_i). A configuration in frame k is an unordered set of 0 or more cylinders $D_k \in \left\{\emptyset \cup C \cup C^2 \cup C^3 \cup \ldots\right\}$. To define a prior distribution over configurations, we restrict location (x, y) to range over a bounded subset W of $\mathbb{R}^2$ and to be distributed with respect to the homogeneous unit intensity Poisson process on W. Cylinder height h is uniformly distributed over a discrete set of heights, independent of location. Prior distribution $P(D_k)$ is therefore a so-called *marked point process* that couples a stochastic point process over the ground plane region of interest W with an additional distribution over a space of attributes at each point (*i.e.* height).

We assume that camera calibration information is known such that 3D locations in the scene can be projected into the 2D image plane, and define the detection likelihood function as a marked Gibbs point process

$$P(Z_k|D_k) \propto \exp\{-U(D_k, Z_k)\} \tag{8}$$

where energy function $U(\cdot)$ measures how well the projection of a hypothesized configuration of cylinders D_k agrees with the image observations Z_k. We compute this energy function using a *count image* S_k where each pixel $p \in S_k$ contains a count of how many cylinders project to that image location. With respect to this count image, the energy function contains two terms

$$\begin{aligned} U &= \alpha_1 G_1(S_k, Z_k) + \alpha_2 G_2(S_k) && (9)\\ G_1 &= \sum_{p \in S_k} \delta\left[\min(1, S(p)) \neq F(p)\right] && (10)\\ G_2 &= \sum_{p \in S_k} (S(p) - 1). && (11) \end{aligned}$$

Energy term G_1 penalizes pixels where there is disagreement between the count image and foreground mask over whether the image location is occupied. Energy term G_2 penalizes pixels where multiple people overlap, thereby encouraging solutions to have the smallest number of people that explain the data.

If association links are not needed or only one image is observed, we can search for the MAP estimate over detections with an RJMCMC sampler (see Section 3.5) having $P(\mathcal{D}|\mathcal{Z})$ as its target distribution. Due to the assumption of independence across frames, this is equivalent to estimating detections D_k independently for each frame k. Figure 3 (bottom row) shows such a single-frame detection result. It has been seen in previous work [29,13] that these single-frame solutions can be quite good at determining the number and location of multiple people whose projections partially occlude each other in the image.

3.4 Associations

The representation of association variables $\mathcal{A}$ is inspired by work on globally optimal data association based on network flow [5,23]. Consider the multi-frame detections in a configuration D to be nodes in a multi-stage trellis graph where each stage corresponds to one frame. Define an edge between each pair of detections in adjacent frames of the graph, *i.e.* such that $d_i \in D_k$ and $d_j \in D_{k+1}$ for $k = 1, \ldots, N-1$. Paired with each edge is a binary *association link* variable f_{ij}. A value $f_{ij} = 1$ means d_i is to be linked with d_j to form one segment of a trajectory, whereas $f_{ij} = 0$ means the association link is turned off. With respect to detection d_k all edges f_{*k} are considered to be *incoming* edges, and all variables f_{k*} are called *outgoing* edges. Each detection d_k also has an incoming and outgoing dummy link ϕ allowing it to be the first or last node of a trajectory, or to be marked as a false positive.

Each edge has an associated cost

$$c_{ij} = \begin{cases} \|d_i - d_j\|^2/\sigma^2 + \mathrm{EMD}(h_i, h_j) & ; \|d_i - d_j\| \leq \rho \\ \beta & ; i = \phi \text{ or } j = \phi \\ \infty & ; \text{otherwise} \end{cases}$$

combining distance information with color similarity measured by Earth Mover's Distance on color histograms h_i and h_j extracted from the image projection of each detection. Link variables with infinite cost can never be turned on and their edges can be excluded from the graph. Parameter ρ is a distance gating threshold, set to the maximum distance a person can travel from one sample frame to the next, σ^2 determines how much small displacements should be favored over larger ones, and β is a penalty for missed detections, which should be set at least as large as the largest gated detection cost would be, *e.g.* $\beta > \rho^2/\sigma^2$.

The likelihood over association variables is defined in Gibbs form as

$$P(A|D, Z) \propto \exp\{-V(A, D, Z)\} \tag{12}$$

where $V(\cdot)$ is a linear function of the association variables $\mathcal{A} = \{f_{ij}\}$:

$$V = \alpha_3 G_3(A, D, Z) \tag{13}$$

$$G_3 = \sum_{f_{ij} \in \mathcal{A}} c_{ij} f_{ij}. \tag{14}$$

Our goal is to choose binary values for variables f_{ij} that minimize the sum of costs in G_3 subject to constraints that exactly one incoming link and one outgoing link to and from each detection is set to 1, and all others to 0. If we connect all incoming dummy links to a distinguished *source* node, and outgoing dummy links to a *sink* node, this minimization can be addressed within a mincost network flow framework (e.g. [28]). However, the constraint that exactly one incoming and outgoing link are turned on makes the problem more natural to view as a multi-dimensional assignment problem [11], which can be be solved efficiently using an algorithm due to Shaffique[25].

Algorithm 1. HYBRID RJMCMC for maximizing $P(\mathcal{A},\mathcal{D}|\mathcal{Z})$

```
Input: $Z$, ITERMAX
Output: $A^*, D^*$

Initialize $D_c$ and $A_c$. Let $A^*, D^* = A_c, D_c$.
for t = 1 to ITERMAX
    choose a stochastic move and propose $D'$
    compute $A' = A(D')$ to maximize $P(\mathcal{A}|D', Z)$
    compute acceptance ratio $\alpha((D_c, A_c) \rightarrow (D', A'))$
    sample $u \sim U(0,1)$
    if $\log(u) < \log(\alpha((D_c, A_c) \rightarrow (D', A')))$
        $D_c, A_c = D', A'$
    if $P(A_c, D_c|Z) > P(A^*, D^*|Z)$
        $A^*, D^* = A_c, D_c$
end
```

3.5 Optimization

We optimize over association and detection variables in $P(\mathcal{A},\mathcal{D}|\mathcal{Z})$ by using a hybrid RJMCMC algorithm that samples over multi-frame detection configurations $\mathcal{D}$, interleaved with deterministic computation of multi-frame data association variables $\mathcal{A}$ with respect to each proposed set of detections (see Algorithm 1). Given a current state (D_c, A_c), the algorithm proposes a new state (D', A') by randomly perturbing the detection configuration $D_c \rightarrow D'$ and then deterministically computing the set of associations A' that maximize $P(A, D'|Z)$. This new state (D', A') is then accepted or rejected according to the Metropolis-Hastings-Green (MHG) ratio $\alpha((D_c, A_c) \rightarrow (D', A'))$ [15]. More details follow.

The stochastic moves used for proposing a transition $D_c \rightarrow D'$ are:
Birth: Choose a frame uniformly at random. Add a new detection to the unordered configuration with location (x_i, y_i) chosen by sampling from the proposal map for that frame and height h_i chosen from a discrete set of height options.
Death: Choose a frame uniformly at random. If there are no detections currently in that frame, no transition occurs. Otherwise, choose a detection uniformly at random and remove it from the configuration for that frame.
Diffusion: Choose a frame uniformly at random. If there are no detections currently in that frame, no transition occurs. Otherwise, choose a detection uniformly at random and perturb its (x_i, y_i) location to $(x_i + dx, y_i + dy)$ with $dx \sim U(-\Delta x, +\Delta x)$ and $dy \sim U(-\Delta y, +\Delta y)$. If the new location is outside the region of interest, no transition occurs. Also choose a new height h_i uniformly at random from the discrete set of height options.

Once a detection configuration D' is proposed, we seek an optimal multi-frame assignment A' to maximize $P(\mathcal{A}|D', Z)$. This is computed by a deterministic function $A' = A(D')$, leveraging the fact that the globally optimal solution to the multidimensional assignment problem of Section 3.4 can be found in strong polynomial time [25]. We prefer the multidimensional assignment framework rather than classical network flow because we want to explicitly penalize false positive detections, not ignore them (in MDA every detection must be explained; in network flow, false positives do not contribute to the cost of the solution if no flow is routed through them). It is important to have these penalties as feedback to encourage the exploration of detection configurations having fewer false positives and missed detections.

All stochastic proposal moves are local updates of detections only, and have a dimension matching Jacobian of 1, so the MHG acceptance ratio [15] reduces to the Metropolis-Hastings ratio:

$$\alpha((D_c, A_c) \to (D', A')) = \min\left(1, \frac{P(D', A'|Z)}{P(D_c, A_c|Z)} \frac{Q(D_c \to D')}{Q(D' \to D_c)}\right) . \tag{15}$$

In this equation, $Q(a \to b)$ is the probability of proposing detection configuration b from the current configuration a, which is very easy to compute in all cases due to the highly localized effects of birth, death and diffusion moves.

3.6 Justification of Correctness

It is widely known that including deterministic moves in an MCMC sampler is dangerous because the chain may become non-ergodic and violate the detailed balance conditions that ensure a correct sampler [18]. Indeed, if our goal was to generate samples (D_c, A_c) representative of the joint distribution $P(\mathcal{A}, \mathcal{D}|\mathcal{Z})$ to make statistical inferences, such as computing expected values, our algorithm above would not be a correct sampler. This is because there are regions of joint $\mathcal{A}, \mathcal{D}$ space that have nonzero probability under $P(\mathcal{A}, \mathcal{D}|\mathcal{Z})$ yet have zero probability of being transitioned to, since the deterministic solution $A = A(D)$ does not maintain any diversity of associations for a given detection configuration. Referring back to Fig. 2, note how only a single "point" along each line of constant $\mathcal{D}$ is ever generated by the sampler.

However, we are using MCMC not for statistical inferencing on $P(\mathcal{A}, \mathcal{D}|\mathcal{Z})$ but to guide search for its global mode. Recall that our sampler can be interpreted as searching for D^* from the max-marginal distribution $\Psi(\mathcal{D})$, and computing $A^* = A(D^*)$ deterministically from that. We therefore should be able to find the global mode A^*, D^* if our algorithm is a correct sampler over $\Psi(\mathcal{D})$. To prove this correctness, first note that $A(D)$ is strictly positive for any argument D, since assigning every detection as a false positive is always an option, and yields a positive value. As a distribution in Gibbs form, $P(\mathcal{D}|\mathcal{Z})$ is also strictly positive for any configuration with countable number of detections. Therefore, any proposed configuration of detections D has a non-zero probability of being accepted. It suffices then to show that RJMCMC with the moves described earlier yields a sampler having stationary distribution $\Psi(\mathcal{D})$. The proof follows Appendix B of van Lieshout [11]. Specifically, the chain is positive recurrent and irreducible with respect to the null configuration of 0 detections in any frame, since any configuration can be transformed with positive probability to the null configuration by a finite series of death moves, and conversely any configuration can be recovered with positive probability by a finite series of birth moves.

Furthermore, there is a positive probability of staying in the null configuration for one or more time steps (for example, if a death move is proposed), and therefore the chain is also aperiodic. These properties are sufficient to ensure that target distribution $\Psi(\mathcal{D})$ is the unique stationary distribution of the chain.

4 Evaluation

In this section we present a proof of concept that the stochastic/deterministic sampler presented above works in practice. We evaluate our method on one in-house sequence and two publicly available video sequences. All were captured from stationary, calibrated cameras, allowing us to estimate object locations and trajectories in a metric ground-plane coordinate system.

Test Sequences: 1) The **Doohan** sequence is a short 20 second clip from an NCAA college basketball game recorded at 25fps and an image resolution of 1920×1456. All 10 players plus 2 referees are visible in the playing area through the whole sequence. Tracking of players is challenging due to their rapid and erratic motion, close proximity, and similar appearance. Ground truth locations were estimated by hand in a floor-plane coordinate system. 2) **APIDIS** sequence is a one minute video from the public APIDIS dataset[2]. It shares the same player tracking challenges as Doohan, but in addition players leave and reenter the field of view and extreme lighting causes saturated regions, long shadows, and poor color quality. The APIDIS dataset has been popular for testing multi-view volumetric tracking approaches [1,3,24]; however, we are interested in evaluating single-view tracking and only use camera 6, which views the right half of the court. Ground truth locations in the floor plane that were annotated every 1 second are distributed with the full APIDIS dataset. 3) The **Oxford Town-Centre** sequence [4] shows pedestrians walking along a busy street, recorded at 25fps with a resolution of 1920×1080[3]. The pedestrian paths are mostly smooth with constant velocity, however there are partial occlusions by signs and benches, and additional objects such as bicycles and strollers appear. This dataset only has meaningful annotations for head locations in the image plane. However, by assuming a constant height of 1.8 meters we approximate the corresponding ground location for each person. This leads to a bias in the "ground truth" for people who are not that tall; particularly noticeable for children in strollers.

Evaluation Metrics: We evaluate both detection and tracking results using the popular CLEAR MOT metrics [6]. MOTP is a measure of geometric accuracy of detections, and is measured for these sequences as distance in the ground plane. MOTA evaluates data association accuracy by penalizing ID swaps, false positives and missed detections along a trajectory. Also reported are precision, recall and average track length. In the online supplemental material we report all intermediate numbers used to compute these measures (e.g. TP, FN, FP, ID swaps),

[2] `http://www.apidis.org/Dataset/`

[3] `http://www.robots.ox.ac.uk/ActiveVision/Publications/benfold_reid_cvpr2011/`

and present additional evaluations with respect to 2D bounding box overlap rather than ground plane distance, evaluation of the effects of color appearance on algorithm performance, and measurement of performance improvement as the number of iterations of the MCMC algorithm increases.

Baseline Algorithms: For comparison, we developed four single-view baseline algorithms for multi-target detection and tracking, generated as the cross-product of two kinds of detectors and two kinds of data association. For detectors, **SF** is a single-frame version of our MCMC detection algorithm (Section 3.3), run on each frame independently, while **POM** is the probabilistic occupancy map detector of [12] run in single-frame mode using the same foreground mask used by our detector. For data association, **Oneshot** applies the deterministic multi-dimensional assignment algorithm [25] that we use solve for associations given a set of detections (Section 3.5), while **DC** applies the discrete-continuous linking and smoothing algorithm of [2] to a given set of detections. All four baseline algorithms (SF-Oneshot, SF-DC, POM-Oneshot, POM-DC) are non iterative, performing a single round of detection followed by a single round of data association, unlike our full algorithm, **HybridFull**, which iteratively performs a stochastic search through the space of multi-frame detection configurations while deterministically solving for the best data association for each configuration.

Results

Table 1 presents the quantitative evaluation results for each tested algorithm on each of the three sequences. Generally, our proposed stochastic/deterministic algorithm achieves the best performance across all measures and all sequences.

The discrete-continuous optimization algorithm [2] incorporates a trajectory smoothing stage, which appears to hinder performance when the temporal sampling is sparse. Additionally, this algorithm does not use appearance information, making it much more difficult to deduce the correct tracking when two objects are in close proximity. However, a clear trend among the algorithms is the longer track length produced by our method. The method of [2] tends to fragment long single-object trajectories into reliable, but short, tracks.

Our 'OneShot' algorithm is essentially the method of [25], but without the ability to infer associations across multi-frame gaps (*i.e.* it is not allowed to compensate for missed detections). In all sequences, HybridFull outperforms SF-OneShot indicating that the MCMC sampler was able to find a better set of detections than the initial solution. This reinforces the point that prematurely fixing the set of detections imposes a burden (e.g. gaps and false detections) that efficient polynomial time data association algorithms cannot overcome.

Figure 4 illustrates the regularizing effect that simultaneous estimation of data association has on estimated detections. Although there are no prior terms encouraging the number of detections in adjacent frames to be similar, the likelihood function $P(\mathcal{A}|\mathcal{D},\mathcal{Z})$ for assignment variables contains penalty terms for unassigned detections, indirectly penalizing configurations having different numbers of detections in each frame. As a result, gaps are filled in and short trajectory fragments are linked together into longer, full trajectories.

Table 1. Quantitative evaluation on the Doohan (top), APIDIS (middle), and Oxford Towncentre (bottom) datasets. The match threshold for CLEAR MOT measures is 1 meter, applied in the ground plane. Lower values are better for MOTP; higher values are better for all other scores. Avglen is computed as (average detected path length / average ground truth path length) * 100.

Doohan sequence					
Algorithm	**MOTP(m)**	**MOTA(-)**	**Prec(%)**	**Recall(%)**	**AvgLen(%)**
SF-OneShot	0.27	86.38	96.14	96.14	33.40
SF-DC	**0.23**	64.63	**97.01**	85.77	8.40
HybridFull(ours)	0.26	**90.24**	96.74	**96.54**	**88.20**
APIDIS sequence					
Algorithm	**MOTP(m)**	**MOTA(-)**	**Prec(%)**	**Recall(%)**	**AvgLen(%)**
SF-OneShot	0.37	45.40	81.38	77.87	23.96
SF-DC	0.43	41.09	78.64	66.67	35.14
POM-OneShot	0.39	17.53	66.01	77.01	11.82
POM-DC	0.49	30.75	71.79	65.80	33.23
HybridFull(ours)	**0.34**	**62.64**	**85.50**	**81.32**	**60.70**
TownCentre sequence					
Algorithm	**MOTP(m)**	**MOTA(-)**	**Prec(%)**	**Recall(%)**	**AvgLen(%)**
SF-OneShot	0.46	29.88	65.06	72.59	32.90
SF-DC	0.59	-17.79	44.67	58.72	53.09
POM-OneShot	**0.40**	19.35	60.57	70.39	20.20
POM-DC	0.61	-5.51	49.19	59.38	59.61
HybridFull(ours)	0.45	**41.32**	**70.45**	**73.84**	**69.06**

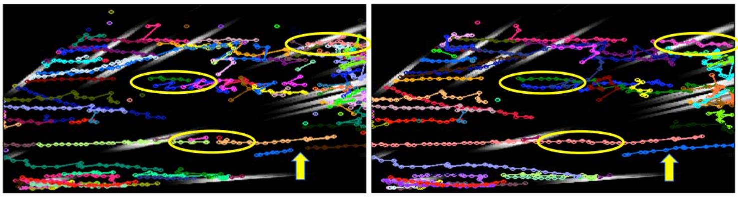

Fig. 4. Top: Initial ground plane trajectories at the start of MCMC processing of the Oxford Town Centre Dataset. Bottom: Final trajectories after 10000 iterations. Several places have been highlighted to illustrate improvements due to gap spanning and trajectory smoothing.

Figure 5 shows sample tracking results from each sequence. Consistency of estimated identity is indicated by bounding boxes of the same color on the same person over time. Videos suitable for qualitative assessment of the results across all frames are available in the online supplemental material.

5 Summary

Traditional tracking-by-detection methods must incorporate complex data association models to handle missed detections and false detections. Our approach, on the other hand, continually explores the set of multi-frame detections and uses a simple data association model for which an optimal solution can be computed efficiently. The burden of dealing with missed and false detections is now handled by the search over multi-frame detections, relying on the power of MCMC

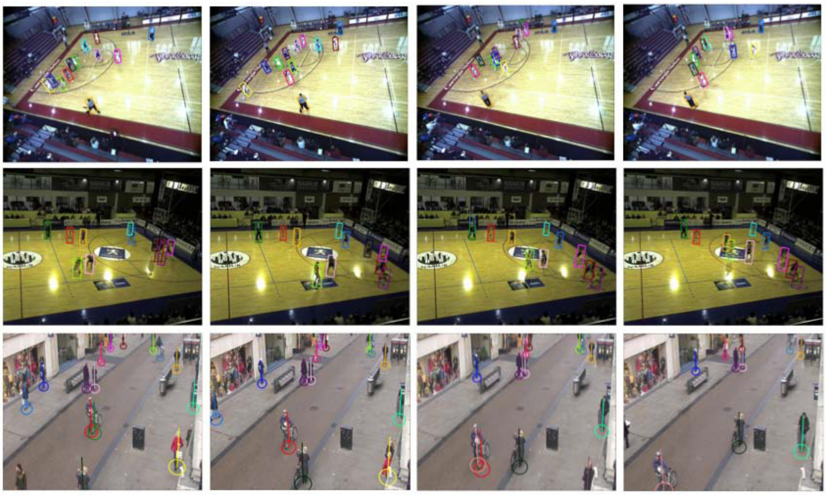

Fig. 5. Sample output frames from the three test sequences. Top: Doohan sequence, Middle: APIDIS camera 6, Bottom: Oxford TownCentre. Color is used to indicate identity across frames.

sampling to produce a near-optimal set of detections. Unlike typical MCMC tracking, which must hypothesize both detections and their associations (and therefore propose complex multi-track moves such as split, merge or swap), our MCMC approach only proposes detections using simple local update moves of birth, death and diffusion while the corresponding associations are computed in a deterministic fashion. As we have shown, incorporating this deterministic aspect into the random search does not jeopardize the necessary conditions of MCMC to have a unique stationary distribution corresponding to our desired target distribution $\Psi(\mathcal{D})$. Additionally, our experiments show how re-estimating the set of multi-frame detections leads to significant improvements in tracking performance.

In future work, the approach in this paper could be generalized in several ways. Three specific ideas are: 1) use proposal moves that refer to the current trajectory estimates when hypothesizing new detections, for example to favor proposing detections that extend a partial track; 2) to use an appearance-based pedestrian detector confidence map in the image to propose pedestrian locations and evaluate their likelihood; and 3) to use a deterministic data association approach that is not strictly guaranteed to yield a global optimum, but that would allow use of more expressive objective functions that include terms of higher-order than the pairwise terms used in network flow / MDA.

References

1. Alahi, A., Jacques, L., Boursier, Y., Vandergheynst, P.: Sparsity driven people localization with a heterogeneous network of cameras. Journal of Mathematical Imaging and Vision 41(1-2), 39–58 (2011)
2. Andriyenko, A., Schindler, K., Roth, S.: Discrete-continuous optimization for multi-target tracking. In: CVPR (2012)
3. Ben Shitrit, H., Berclaz, J., Fleuret, G., Fua, P.: Multi-Commodity Network Flow for Tracking Multiple People. PAMI (2013)
4. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: CVPR (2011)
5. Berclaz, J., Turetken, E., Fleuret, F., Fua, P.: Multiple Object Tracking using K-Shortest Paths Optimization. PAMI 33(9), 1806–1819 (2011)
6. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The CLEAR MOT metrics. EURASIP Journal on Image and Video Processing, Special Issue on Video Tracking in Complex Scenes for Surveillance Applications 2008, article ID 246309 (May 2008)
7. Brau, E., Barnard, K., Palanivelu, R., Dunatunga, D., Tsukamoto, T., Lee, P.: A generative statistical model for tracking multiple smooth trajectories. In: CVPR, pp. 1137–1144 (2011)
8. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multiperson tracking-by-detection from a single, uncalibrated camera. PAMI 33(9), 1820–1833 (2011)
9. Brendel, W., Amer, M., Todorovic, S.: Multiobject tracking as maximum weight independent set. In: CVPR (2011)
10. Carr, P., Sheikh, Y., Matthews, I.: Monocular object detection using 3D geometric primitives. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 864–878. Springer, Heidelberg (2012)
11. Collins, R.: Multitarget data association with higher-order motion models. In: CVPR (2012)
12. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multi-Camera People Tracking with a Probabilistic Occupancy Map. PAMI 30(2), 267–282 (2008)
13. Ge, W., Collins, R.: Marked point processes for crowd counting. In: CVPR. pp. 2913–2920 (2009)
14. Ge, W., Collins, R.: Multi-target data association by tracklets with unsupervised parameter estimation. In: BMVC (2008)
15. Green, P.: Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82(4), 711–732 (1995)
16. Huang, C., Wu, B., Nevatia, R.: Robust object tracking by hierarchical association of detection responses. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 788–801. Springer, Heidelberg (2008)
17. Khan, Z., Balch, T., Dellaert, F.: Mcmc data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements. PAMI 28(12), 1960–1972 (2006)
18. Kim, W., Lee, K.: Markov chain monte carlo combined with deterministic methods for markov random field optimization. In: CVPR (2009)
19. Leibe, B., Schindler, K., Gool, L.J.V.: Coupled detection and trajectory estimation for multi-object tracking. In: ICCV, pp. 1–8 (2007)
20. van Lieshout, M.: Depth map calculation for a variable number of moving objects using markov sequential object processes. PAMI 30(7), 1308–1312 (2008)

21. Milan, A., Roth, S., Schindler, K.: Continuous energy minimization for multitarget tracking. PAMI 36(1), 58–72 (2014)
22. Oh, S., Russell, S., Sastry, S.: Markov chain monte carlo data association for multi-target tracking. IEEE Transactions on Automatic Control 54(3), 481–497 (2009)
23. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR (2011)
24. Possegger, H., Sternig, S., Mauthner, T., Roth, P., Bischof, H.: Robust real-time tracking of multiple objects by volumetric mass densities. In: CVPR, pp. 2395–2402 (2013)
25. Shafique, K., Shah, M.: A noniterative greedy algorithm for multiframe point correspondence. PAMI 27(1), 51–65 (2005)
26. Wu, Z., Thangali, A., Sclaroff, S., Betke, M.: Coupling detection and data association for multiple object tracking. In: CVPR, pp. 1–8. Rhode Island (June 2012)
27. Roshan Zamir, A., Dehghan, A., Shah, M.: Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 343–356. Springer, Heidelberg (2012)
28. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR (2008)
29. Zhao, T., Nevatia, R.: Bayesian human segmentation in crowded situations. In: CVPR (2003)