Distributed Data Association and Filtering for Multiple Target Tracking

Ting Yu† Ying Wu‡ Nils O. Krahnstoever† Peter H. Tu†
yut@research.ge.com yingwu@ece.northwestern.edu {krahnsto,tu}@research.ge.com

† Visualization and Computer Vision Lab, GE Global Research, Niskayuna, NY 12309
‡ Dept. of Electrical Engineering and Computer Sciences, Northwestern University, Evanston, IL 60208

Abstract

This paper presents a novel distributed framework for multi-target tracking with an efficient data association computation. A decentralized representation of trackers' motion and association variables is adopted. Considering the interleaved nature of data association and tracker filtering, the multi-target tracking is formulated as a missing data problem, and the solution is found by the proposed variational EM algorithm. We analytically show that 1) the posteriori distributions of trackers' motions (the real interests in terms of tracking applications) can be effectively computed in the E-step of the EM iterations, and 2) the solution of trackers' association variables can be pursued under a derived graph-based discrete optimization formulation, thus efficiently estimated in the M-step by the recently emerging graph optimization algorithms. The proposed approach is very general such that sophisticated data association priori and likelihood function can be easily incorporated. This general framework is tested with both simulation data and real world surveillance video. The reported qualitative and quantitative studies verify the effectiveness and low computational cost of the algorithm.

1. Introduction

Originally from the radar-tracking literature [2,5], multi-target tracking in video has been actively pursued, while stays as one of the most challenging topics in computer vision [6,8,10,12,19] for decades. Multi-target tracking deals with the state estimation of an unknown number of moving targets, a theoretical problem with tremendous value in real-world applications, such as people tracking in video surveillance, sports video annotation, and vision-based HCI.

Fundamentally different from single target tracking, multi-target tracking algorithm requires a complex data association logic to partition the detected measurements to each individual data source, and establish their correspondence with the maintained trackers. This implies two important processes that critically decide the success of a multi-target tracking algorithm, 1) tracker-measurement associations and 2) tracker filtering, which are in essence two interleaved properties. On one hand, we have to know the trackers' states to obtain good estimation of the measurement-tracker associations; on the other hand, we also need to know the accurate measurement-tracker associations to correctly filter the trackers.

Common approaches to tackle this problem take a centralized representation of a joint association vector, which is then estimated either by exhaustive enumerations, such as the joint probabilistic data association (JPDA) filter [2, 12] or probabilistic Monte Carlo optimization [5]. For example, one of the primary work on multi-target tracking in vision is studied by [12], where a JPDA filter is leveraged to compute all joint association events between the measurements (object detections) and trackers. To constrain the number of events to a manageable level, gating technique [2], i.e., the early pruning of very unlikely association events, has to be applied to reduce the computations.

Living in the joint state space of multiple targets, sampling-based approaches have also been proposed to model the joint likelihood function, thus estimating the combined state of all targets directly [6, 8, 10, 19]. Without resorting to explicitly compute the data association, these approaches demonstrate the capabilities of tracking multiple targets when complex motions are present. Unfortunately, due to the centralized nature of the joint state representation, the complexity of these approaches grows exponentially in the number of targets to be tracked. [18] proposed to work in a decentralized state space to tackle multi-target tracking. Though the primary focus there is to address the "coalescence" problem in a distributed fashion, their method is lack of a principled way to address the problem of identity switch (data association).

Motivated to solve the computational challenges of the

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

existing methods due to the centralized representation on either data association vector, state variable or both, we propose a novel distributed framework for multi-target tracking with an efficient computation on data association and state estimation. A decentralized representation on both trackers' motion states and association variables is adopted. To model the interleaved nature of data association and tracker filtering, the multi-target tracking is formulated as a missing data problem, and the solution is found by the proposed variational EM algorithm. Rigorous derivation of the algorithm arrives an interesting solution that analytically shows 1) the posteriori distributions of trackers' motions (the real interests in terms of tracking applications) can be effectively computed in the E-step of the variational EM iterations, and 2) the solution of trackers' association variables can be pursued under a derived graph-based discrete optimization formulation, and can be efficiently estimated in the M-step by the recently emerging graph optimization algorithms, such as max-product belief propagation and its advances [16].

There are many benefits brought by the proposed approach. Firstly, computationally wise speaking, the involved computations for obtaining the optimal data association and motion filtering are greatly reduced compared to the centralized approaches, such as [2, 12, 14], which makes the algorithm well scalable to track large number of targets. Secondly, the proposed algorithm enjoys tremendous generalization flexibilities such that sophisticated data association modeling and likelihood function can be easily incorporated. Thirdly, the parallel nature of the graph-based data association optimization makes it particularly suitable to run the algorithm in a distributed computational architecture, a very appealing feature to apply the algorithm in real-world applications.

Unified under this general framework, both simulation data solved by a Kalman Filter and real-world surveillance videos tackled by a part-based Particle Filter are presented, where both the Kalman Filter and the Particle Filter are embedded into the proposed variational EM iterations. The reported qualitative and quantitative studies verify the effectiveness and low computational cost of the algorithm.

2. Problem Formulation

To ease the exposition, let us firstly consider a basic case of multi-target tracking problem. More complicated modeling, such as complex data association priori and part-based object representation, as well as some practical issues, such as tracker initialization and termination, will be addressed in later sections.

2.1. Missing Data Formulation

Denote the m_t detected measurements at the current frame t by Z_t, which is composed with $Z_t = \{z_{1,t}, \ldots, z_{m_t,t}\}$. The measurement data collected over frames is depicted by Z^t, and $Z^t = \{Z_1, \ldots, Z_t\}$.

We take a distributed representation for the set of M trackers. Each tracker i, where i represents the tracker identifier and $i \in \{1, \ldots, M\}$, has two unknown variables to be estimated, $\{a_{i,t}, x_{i,t}\}$. $a_{i,t}$ denotes the association variable of tracker i, and takes values from the discrete set $\{0, 1, \ldots, m_t\}$. The tracker i can associate itself with every possible measurement $z_{a_{i,t},t}$ from Z_t, or associate with nothing $a_{i,t} = 0$, indicating the missing detection of the target i or target leaving (disappearing) from the field. The motion state of each tracker i is described by $x_{i,t}$.

$a_t = \{a_{1,t}, \ldots, a_{M,t}\}$, $x_t = \{x_{1,t}, \ldots, x_{M,t}\}$ collect the associations and motion states of all maintained trackers. In essence, multi-target tracking algorithm deals with the problem of estimating the posteriori probability $p(x_t, a_t|Z^t)$. However, due to the heavily interleaved nature of (a_t, x_t), jointly estimating $p(x_t, a_t|Z^t)$ is very challenging. Instead, we propose to estimate the marginal posteriori over one variable and treat the other one as hidden under the missing data formulation, and solve the problem by a variational EM algorithm [7].

Between x_t and a_t, which one should be chosen as the missing variable? Though previous literature usually treats association variables a_t as missing [13, 15], we believe better justifications existing to support x_t as missing. By treating x_t as missing, firstly, we can immediately have a continuously increased estimation of the probabilistic distribution over x_t in the E-step of the EM iterations, which is actually the real interest of target tracking; secondly, as we shall demonstrate later that in the M-step the point estimate of the association variable a_t can be effectively and efficiently optimized by the recently emerging graph-based optimization techniques, such as multi-way graph cut algorithm [3] and max-product belief propagation algorithm [16].

We thus formulate the multi-target tracking as a maximum a posteriori (MAP) estimation problem of the data association variables a_t as follows

$$a_t^* = \arg\max_{a_t} E(a_t) = \arg\max_{a_t} \log p(a_t|Z^t) \tag{1}$$

Such a marginal posteriori $p(a_t|Z^t)$ is obtained by integrating over the unknown motions of the trackers, x_t (missing data), i.e.,

$$a_t^* = \arg\max_{a_t} \log \int_{x_t} p(a_t, x_t|Z^t) dx_t \tag{2}$$

From the Jensen's inequality, a $Q(x_t)$ function can be introduced to break the above logarithm of the integral into a

more manageable lower bound energy function as

$$\begin{aligned} a_t^* &= \arg\max_{a_t} \log \int_{x_t} Q(x_t) \frac{p(a_t, x_t|Z^t)}{Q(x_t)} dx_t \\ &\geq \arg\max_{a_t, Q(x_t)} \int_{x_t} Q(x_t) \log \frac{p(a_t, x_t|Z^t)}{Q(x_t)} dx_t \\ &= \arg\max_{a_t, Q(x_t)} \tilde{E}(a_t, Q(x_t)) \end{aligned} \tag{3}$$

where the equality holds only when the optimal association a_t^* is found and $Q(x_t) = p(a_t^*, x_t|Z^t)$. Maximizing the original objective function $E(a_t)$ can be achieved by iteratively maximizing the lower bound function $\tilde{E}(a_t, Q(x_t))$ over its two unknown properties, a_t and $Q(x_t)$ [7].

2.2. Equivalence To the Variational Analysis

The above missing data formulation can be explained from the variational analysis point of view [7]. In fact, if we can assume that the correct data association variable a_t^* is available as a known model parameter, the optimization of $\tilde{E}(a_t^*, Q(x_t))$ is then only carried out over the unknown distribution $Q(x_t)$. It is equivalent to minimize the Kullback-Leibler (KL) divergence between $Q(x_t)$ and posteriori distribution $p(x_t|a_t^*, Z^t)$ because

$$\begin{aligned} Q^*(x_t) &= \arg\max_{Q(x_t)} \int_{x_t} Q(x_t) \log \frac{p(a_t^*, x_t|Z^t)}{Q(x_t)} dx_t \\ &= \arg\max_{Q(x_t)} \int_{x_t} Q(x_t) \log \frac{p(a_t^*, x_t|Z^t)}{Q(x_t)} dx_t - \log p(a_t^*|Z^t) \\ &= \arg\min_{Q(x_t)} KL(Q(x_t) \parallel p(x_t|a_t^*, Z^t)) \end{aligned} \tag{4}$$

where since a_t^* is known, $\log p(a_t^*|Z^t)$ becomes a constant, thus when plugged into the equation, does not change the optimality condition. $Q(x_t)$ is called a variational distribution, and is an approximate distribution to $p(x_t|a_t^*, Z^t)$. Thus the problem becomes to find a best approximation $Q(x_t)$ to minimize this KL-divergence. In variational analysis, if we choose a fully factorized form of $Q(x_t)$, i.e.,

$$Q(x_t) = \prod_i^M Q_i(x_{i,t}) \tag{5}$$

we end up with mean field approximation [7], where each factorial $Q_i(x_{i,t})$ is to approximate the unknown marginal probabilities $p(x_{i,t}|Z^t)$. In reality, since we do not know the optimal association a_t^* before hand, we have to estimate both $Q(x_t)$ and a_t simultaneously. Such a process is called as variational EM iterations in literature [7].

2.3. Model Expansion

Given $\tilde{E}(a_t, Q(x_t))$ in Eq. 3, further expansion gives:

$$\begin{aligned} &\arg\max_{a_t, Q(x_t)} \tilde{E}(a_t, Q(x_t)) \\ &= \arg\max_{a_t, Q(x_t)} \int_{x_t} Q(x_t) \log p(a_t, x_t|Z^t) dx_t + H(Q(x_t)) \\ &= \arg\max_{a_t, Q(x_t)} \int_{x_t} Q(x_t) \log p(a_t, x_t|Z^t) dx_t + H(Q(x_t)) + \log p(Z_t|Z^{t-1}) \\ &= \arg\max_{a_t, Q(x_t)} \int_{x_t} Q(x_t) \log p(a_t, x_t, Z_t|Z^{t-1}) dx_t + H(Q(x_t)) \end{aligned} \tag{6}$$

where $H(Q(x_t))$ is the entropy of $Q(x_t)$, and $p(Z_t|Z^{t-1})$ is an added constant once the measurement data is given and does not change the optimal condition. However, such a plug-in does bring an interesting chain rule expansion, i.e.,

$$p(a_t, x_t, Z_t|Z^{t-1}) = p(x_t|Z^{t-1}) p(a_t|x_t, Z^{t-1}) p(Z_t|a_t, x_t, Z^{t-1}) \tag{7}$$

thus to model the problem, we need reasonable models for each of the above three distributions 1) prediction probability $p(x_t|Z^{t-1})$ 2.3.1, 2) priori probability of the association variable $p(a_t|x_t, Z^{t-1})$ 2.3.2, and 3) likelihood model $p(Z_t|a_t, x_t, Z^{t-1})$ 2.3.3. We explain them one by one later.

With the reasonable Markovian assumption, we can simplify $p(a_t|x_t, Z^{t-1}) = p(a_t|x_t)$ and $p(Z_t|a_t, x_t, Z^{t-1}) = p(Z_t|a_t, x_t)$. Thus, the maximization problem becomes

$$\begin{aligned} &\max_{a_t, Q(x_t)} \tilde{E}(a_t, Q(x_t)) = \max_{a_t, Q(x_t)} H(Q(x_t)) \\ &+ \int_{x_t} Q(x_t) \log[p(x_t|Z^{t-1}) p(a_t|x_t) p(Z_t|a_t, x_t)] dx_t \end{aligned} \tag{8}$$

2.3.1 Motion Prediction, $p(x_t|Z^{t-1})$

$p(x_t|Z^{t-1})$ is the motion prediction model of the trackers

$$p(x_t|Z^{t-1}) = \int_{x_{t-1}} p(x_t|x_{t-1}) p(x_{t-1}|Z^{t-1}) dx_{t-1} \tag{9}$$

This joint motion posteriori $p(x_{t-1}|Z^{t-1})$ can be suitably approximated via the product of its marginal components $p(x_{i,t-1}|Z^{t-1})$, i.e., $p(x_{t-1}|Z^{t-1}) \approx \prod_{i=1}^M p(x_{i,t-1}|Z^{t-1})$. Recall that the optimal Q-function $Q_i^*(x_{i,t-1})$ for tracker i from frame $t-1$ is a good approximation of the tracker's motion posteriori $p(x_{i,t-1}|Z^{t-1})$, and also employ an independent dynamics models $p(x_t|x_{t-1}) = \prod_{i=1}^M p(x_{i,t}|x_{i,t-1})$, the joint motion prediction term $p(x_t|Z^{t-1})$ can then be simplified as the following factorized form

$$p(x_t|Z^{t-1}) \approx \prod_{i=1}^M \int_{x_{i,t-1}} p(x_{i,t}|x_{i,t-1}) Q_i^*(x_{i,t-1}) dx_{i,t-1} \tag{10}$$

2.3.2 Association Priori, $p(a_t|x_t)$

Conditioned on the motions x_t, $p(a_t|x_t)$ is the priori probability of the association variable $a_t = \{a_{1,t}, \ldots, a_{M,t}\}$. It imposes constraints on preventing infeasible data association behaviors, such as the situation that two trackers associate themselves to a single measurement. With a proper modelling of this term, we will end up with a valid partition of the measurement data into different trackers.

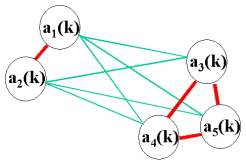


Figure 1. An example of fully connected pair-wise data association constraint.

Unlike the traditional formulation that enumerates every association possibility with a high dimensional joint data association vector [2, 12, 14], we embed the data association constraints (priori) into a graph structure. One type of this graph-based constraint can be formulated as a fully-connected but distributed pair-wise graph as shown in Figure 1, and the corresponding probabilistic model is

$$p(a_t|x_t) = \frac{1}{Z_{x_t}} \prod_{(i,j)\in E} \psi(a_{i,t}, a_{j,t}|x_t) \tag{11}$$

where E denotes the set of neighbors in that we introduce the constraint, and $\psi(a_{i,t}, a_{j,t}|x_t)$ is the pair-wise constraint between $a_{i,t}$ and $a_{j,t}$. Z_{x_t} is a partition function to make $p(a_t|x_t)$ a proper probability distribution. In Figure 1, each circle depicts an association variable of a tracker, and the edges connecting them represent the existence of pair-wise association constraints[1]. A common practice of employing $p(a_t|x_t)$ is to assume its independence from trackers' motions x_t [2, 14], i.e.,

$$p(a_t|x_t) = p(a_t) = \frac{1}{Z} \prod_{(i,j)\in E} \psi(a_{i,t}, a_{j,t}) \tag{12}$$

a pure association priori, where the pair-wise constraint is:

$$\psi(a_{i,t}, a_{j,t}) = \begin{cases} 0, & \{a_{i,t} = a_{j,t} \neq 0\} \\ 1, & \text{otherwise} \end{cases} \tag{13}$$

Remember that both $a_{i,t}$ and $a_{j,t}$ can choose values from the discrete measurement set $\{0, 1, \ldots, m_t\}$, thus the basic idea of $\psi(a_{i,t}, a_{j,t})$ is to explicitly avoid the case that two trackers associate themselves to a single measurement unless it is a missing detection or occlusion event $a_{i,t} = a_{j,t} = 0$. In essence, the association constraint in Eq. 12 enforces to only generate valid association hypotheses a_t.

[1] we intentionally thicken the edge link between $(a_{1,t}, a_{2,t})$ and the ones among $(a_{3,t}, a_{4,t}, a_{5,t})$ to illustrate that the trackers in each of these two groups are spatially closer to each other, thus "gating technique" [2] can also be applied here by removing other weaker links.

2.3.3 Likelihood Model, $p(Z_t|a_t, x_t)$

$p(Z_t|a_t, x_t)$ is the joint likelihood model of the measurement data Z_t, conditioned on (a_t, x_t). Without knowing a_t, the joint likelihood $p(Z_t|x_t)$ can not be factorized, i.e., $p(Z_t|x_t) \neq \prod_{i=1}^{M} p(z_{i,t}|x_{i,t})$. On the contrary, if a_t is provided, we can factorize this joint likelihood model, since we know which measurement data $z_{a_{i,t},t}$ is generated from each tracker $x_{i,t}$, i.e., we have

$$p(Z_t|a_t, x_t) = \prod_{i=1}^{M} p(z_{a_{i,t},t}|x_{i,t}) \tag{14}$$

To generate $p(z_{a_{i,t},t}|x_{i,t})$, firstly we need a reasonable model for the usual case of $p(z_{a_{i,t}\neq 0,t}|x_{i,t})$, i.e., the tracker i is associating with a valid measurement. The exact form of this term relies on the domain knowledge. In Kalman Filtering framework, this measurement model is defined by a Gaussian distribution [2, 12]; while in the visual tracking scenario, $p(z_{a_{i,t}\neq 0,t}|x_{i,t})$ can also incorporate any available visual attributes, such as appearance, shape, etc [6, 10, 17, 19]. Secondly, $p(z_{a_{i,t}=0,t}|x_{i,t})$ needs some special modelling, where $a_{i,t} = 0$ means the target followed by the tracker $x_{i,t}$ is missing detected, occluded, or leaving the scene [14][2]. The value range of this special likelihood is critical, and a proper setting will not only support the existence of a tracker to follow a temporarily occluded target during the occlusion period, but also maintain the tracker to tolerate missing detections. [9, 14] discussed some principled ways of making a proper selection of this value.

2.4. EM Solution

From above, we understand that the motion prediction $p(x_t|Z^{t-1})$, association priori $p(a_t|x_t)$, and likelihood $p(Z_t|a_t, x_t)$ models all take some factorized or distributed forms. By plugging them into the expanded objective in Eq. 8, and also considering the fully-factorized Q-function in Eq. 5, after some manipulations, we have

$$\{a_t^*, Q^*(x_t)\} = \arg\max_{a_t, Q(x_t)} \sum_{(i,j)\in E} \log \psi(a_{i,t}, a_{j,t}) + \sum_{i}^{M} H(Q_i(x_{i,t})) + \sum_{i=1}^{M} \int_{x_{i,t}} Q_i(x_{i,t}) \log[p(x_{i,t}|Z^{t-1}) p(z_{a_{i,t},t}|x_{i,t})] dx_{i,t} - \log Z \tag{15}$$

Based on this objective, our EM solution involves to solve the following two iterative steps, where the exact computational forms are presented in Sections 2.4.1 and 2.4.2:

E-Step: Compute a better $Q'(x_t) = \prod_{i=1}^{M} Q_i'(x_{i,t})$ over trackers motions x_t to maximize $\tilde{E}(a_t, Q(x_t))$.

[2] Though in principle, the proposed framework could assign different association indexes to missing detection, occlusion, and disappearing events, without losing generality, in this paper we do not differentiate these cases and treat them with a single unified event $a_{i,t} = 0$.

M-Step: Find a better association $a_t^{'} = \{a_{1,t}^{'}, \ldots, a_{M,t}^{'}\}$ to maximize $\tilde{E}(a_t, Q^{'}(x_t))$.

2.4.1 E-Step

Take the partial derivative of the objective in Eq. 15 over $Q_i(x_{i,t})$, and enforce the constraint that each $Q_i(x_{i,t})$ must be a valid probabilistic distribution, i.e., $\int_{x_{i,t}} Q_i(x_{i,t}) dx_{i,t} = 1$ we obtain the E-step updating equation for each tracker i

$$Q_i^{'}(x_{i,t}) \propto p(z_{a_{i,t},t}|x_{i,t}) p(x_{i,t}|Z^{t-1}) \tag{16}$$

which has almost the same form of filtering solution for single-target tracking, i.e., the combination of prediction model and likelihood function. The only difference is that the measurement data $z_{a_{i,t},t}$ used to filter $x_{i,t}$ is conditioned on the association variable $a_{i,t}$, which has to be estimated from the following M-step. Note that the above E-step updating is composed with M independent updating equations, where each one is for an individual tracker. It reflects that the E-step computation is indeed distributed.

2.4.2 M-Step

For M-step, we find an updated set of association variables $a_t = \{a_{1,t}, \ldots, a_{M,t}\}$ to increase the objective given the already updated $Q^{'}(x_t)$ from E-step

$$\begin{aligned} a_t^{'} = & \arg\max_{a_t} \sum_{(i,j)\in E} \log \psi(a_{i,t}, a_{j,t}) \\ & + \sum_{i=1}^{M} \int_{x_{i,t}} Q_i^{'}(x_{i,t}) \log p(z_{a_{i,t},t}|x_{i,t}) dx_{i,t} \end{aligned} \tag{17}$$

where all terms unrelated to a_t are removed. Define the following two items $f_{i,j}(a_{i,t}, a_{j,t})$ and $g_i(a_{i,t})$

$$\begin{aligned} & f_{i,j}(a_{i,t}, a_{j,t}) = \psi(a_{i,t}, a_{j,t}) \\ & g_i(a_{i,t}) = \exp\{\int_{x_{i,t}} Q_i^{'}(x_{i,t}) \log p(z_{a_{i,t},t}|x_{i,t}) dx_{i,t}\} \end{aligned} \tag{18}$$

which are the functions of $(a_{i,t}, a_{j,t})$ and $a_{i,t}$ respectively. Also notice that the optimality condition does not change if an $\exp$ operation is applied to an objective function. The objective in Eq. 17 then is further written as

$$a_t^{'} = \arg\max_{a_t} \prod_{(i,j)\in E} f_{i,j}(a_{i,t}, a_{j,t}) \prod_{i=1}^{M} g_i(a_{i,t}) \tag{19}$$

Remind that $\{a_{1,t}, \ldots, a_{M,t}\}$ are from a discrete value set, therefore both $f_{i,j}(a_{i,t}, a_{j,t})$ and $g_i(a_{i,t})$ can be precomputed before the M-step optimization. Eq. 19 refers to one of the most standard forms of optimization problems that can be solved by many emerging algorithms. Considering the involved graph structure of the problem as shown in Figure 1, efficiently approximate solutions exist, such as the multi-way graph cut algorithm [3], the max-product belief propagation (BP) [16]. In our implementation, we use the max-product BP to obtain the optimal solution $a^{'}(k)$ in this arbitrary graph structure. The BP algorithm and its variants are distinguished with their distributed and parallel computational paradigm, thus the derived M-step updating is also bestowed with this distributed computation via the use of max-product BP.

Through above EM analysis, it is clear that both E-step (motion filtering) and M-step (data association) enjoy the distributed and parallel computational nature, which assures us to claim that the proposed algorithm is essentially a distributed solution compared to existing methods. The efficiency and effectiveness of M-step (BP algorithm) were extensively addressed in literature [7,16], and the efficiency of E-step is directly reflected by the individual filtering equation of each tracker.

3. Model Generalizations

3.1. Part-based Object Representation

The description so far considers a simple object model, where we assume a holistic-based object classifier is available to detect the target, such as pedestrians in surveillance application [4]. The holistic-based detector works fine when targets are moving in isolated mode or under minor occlusions. It is, however, not suitable to deal with large mutual occlusions. Though in theory the likelihood model $p(z_{a_{i,t}=0,t}|x_{i,t})$ can support the existence of a tracker during occlusion, this likelihood is more suitable to the situation of full occlusion. Under partial occlusion, it makes more sense to leverage the minor but still visible part information of the occluded target. Therefore, an object detector with part-based representation [17] is better pursued.

Given a K-part decomposition of the object, such as human samples illustrated in Figure 4, where $K = 3$ stands for head-shoulder, torso, and legs [17], K detectors are trained by collecting the training data of each corresponding part. The association variable of a tracker is now formed by K parts, i.e., $a_{i,t} = (a_{i,1,t}, \ldots, , a_{i,K,t})$, where each $a_{i,k,t}, k \in K$ describes an association that assigns a part detection from the corresponding part detector to tracker i. The motion state of the tracker is still $x_{i,t}$. Conditioned on $x_{i,t}$ and association variable $a_{i,t}$, the likelihood function $p(z_{a_{i,t},t}|x_{i,t})$ in Eq. 14 is modelled by the product of its component likelihoods, i.e., $p(z_{a_{i,t},t}|x_{i,t}) = \prod_{k=1}^{K} p(z_{a_{i,k,t},t}|x_{i,t})$. The association priori in Eq. 12 becomes $p(a_t) = \frac{1}{Z} \prod_{(i,j)\in E} \prod_{k=1}^{K} \psi(a_{i,k,t}, a_{j,k,t})$. This part-based model generalization does not change the Q-function computation in Eq. 16 in E-step, while the only extra computations induced are in M-step. Rather than solving one graph optimization defined in Eq. 19, K graph op-

timizations need to be carried out to obtain the optimal part associations $a'_{i,t} = (a'_{i,1,t}, \ldots, a'_{i,K,t})$ simultaneously.

3.2. Depth-based Association Priori

In Section 2.3.2, we mentioned that it is common to assume the association priori as $p(a_t|x_t) = p(a_t)$, i.e., a_t independent of x_t [2, 14]. In general, this pure association priori $p(a_t)$ is effective. In practice, when x_t can benefit more from the scene, we may build a better association priori than the one in Eq. 13 by keeping the dependencies of a_t on x_t. One possible form is $p(a_t|x_t) = \prod_{(i,j)\in E} p(a_{i,t}, a_{j,t}|x_{i,t}, x_{j,t})$. We replace ψ with p and drop the partition function Z_{x_t}, to explicitly note that each pair-wise constraint $p(a_{i,t}, a_{j,t}|x_{i,t}, x_{j,t})$ holds its own normalization.

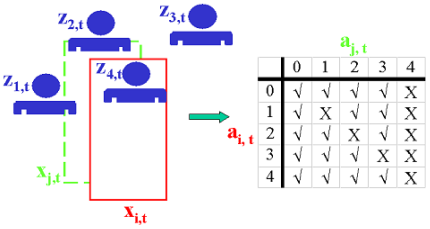


Figure 2. An illustrative example of depth-based pair-wise association constraint.

For example, if scene geometry is available, a tracker can run in $3D$ space. The motion state of each $3D$ tracker contains depth information. Let $x_{i,t} < x_{j,t}$ denote the motion hypothesis that tracker i is closer to the camera than tracker j. Illustrated in Figure 2, two rectangle boxes (solid-red and dash-green) represent a configuration of two trackers i and j with $x_{i,t} < x_{j,t}$, and four detections $Z_t = \{z_{1,t}, \ldots, z_{4,t}\}$ are returned. Conditioned on $x_{i,t} < x_{j,t}$, the four detections firstly are partitioned into, $Z_t^1 = \{z_{1,t}, z_{2,t}, z_{3,t}\}$ and $Z_t^2 = \{z_{4,t}\}$, depending on whether a detection is covered by the projection of the front tracker i. With $x_{i,t} < x_{j,t}$, all valid pair-wise association constraints $p(a_{i,t}, a_{j,t}|x_{i,t}, x_{j,t})$ are listed in the table of Figure 2, where $\surd$ represents an allowable association, while X means not. From the table, we see that besides the common constraint $\{a_{i,t} = a_{j,t} \neq 0\}$, all configurations with $a_{j,t} = 4$ become unacceptable, since the tracker hypothesis says that $x_{i,t} < x_{j,t}$. Such a depth-based association priori will affect the EM solutions on both E-step and M-step. Without describing the details, we simply write down the changed EM iterations

E-step, the iterative solution of Q-function becomes

$$Q'_i(x_{i,t}) \propto \prod_{j\in N(i)} \exp\{\int_{x_{j,t}} Q_j(x_{j,t}) \log p(a'_{i,t}, a'_{j,t}|x_{i,t}, x_{j,t}) dx_{j,t}\}$$
$$\times p(z_{a_{i,t},t}|x_{i,t}) \int_{x_{i,t-1}} p(x_{i,t}|x_{i,t-1}) p(x_{i,t-1}|Z^{t-1}) dx_{i,t-1} \quad (20)$$

where the updating of $Q_i(x_{i,t})$ takes each neighbor's $Q_j(x_{j,t})$ into consideration.

M-step, the M-step objective in Eq. 19 does not change, while the way of pre-computing $f_{i,j}(a_{i,t}, a_{j,t})$ is modified

$$f_{i,j}(a_{i,t}, a_{j,t})$$
$$= \exp\{\int_{x_{i,t}, x_{j,t}} Q'_i(x_{i,t}) Q'_j(x_{j,t}) \log p(a_{i,t}, a_{j,t}|x_{i,t}, x_{j,t}) dx_{i,t} dx_{j,t}\} \quad (21)$$

where an integral evaluated over the motions of pair-wise trackers $(x_{i,t}, x_{j,t})$ are needed to pre-compute $f_{i,j}(a_{i,t}, a_{j,t})$.

Compared to the basic forms of EM steps in Section 2.4, depth-based association priori induces extra integral computations over any pair-wise trackers. Though challenging to compute, the proposed solution is still much more manageable than existing methods that leverage the depth-based scene geometry, where an exhaustive sampling over the joint state space of all trackers is required [6, 10, 19].

4. Experiments

The proposed algorithm is evaluated against both simulation data and real-world surveillance video. Please note for all experiments, a fully-connected pair-wise graph is adopted, i.e., no gating technique [2] is employed to prune the computation. Due to the decentralized nature of the tracker, tracker initialization and termination are very straight-forward. An established tracker i terminates if it is not associating with any valid measurement for n consecutive frames. Any unassociated measurement after the EM iterations will initialize a temporarily new tracker, which will be maintained and confirmed to be an established one only if it survives for m frames.

4.1. Simulation Results

We firstly examine the algorithm on challenging simulation sequences, and compared against a JPDA filter [2]. Simulation provides a controllable setting to have a fair comparison. Because both motion dynamics in Eq. 10 and likelihood model in Eq. 14 are assumed to be Gaussian, Kalman Filter is directly applied here to compute the E-step in Eq. 16, i.e., the Q-function $Q_i(x_{i,t})$ is completely characterized by the state estimate and state covariance in Kalman Filter setting [2]. The integral computation in the term of $g_i(a_{i,t})$, required during M-step optimization, can also be analytically computed by observing $Q'_i(x_{i,t})$ is a Gaussian and $\log p(z_{a_{i,t},t}|x_{i,t})$ is only a quadratic form. Therefore, except the max-product belief propagation involved in the M-step, every remaining step lives in the Kalman Filter world. Two important parameters in the simulation are 1) λ_c, the density parameter of Possion clutter model, which we set $\lambda_c = 5$, and 2) P_d, target detection rate, which is

set to $P_d = 0.9$ in our experiments. Two snapshots of running our algorithm on a simulation sequence to track three targets are shown in Figure 3.

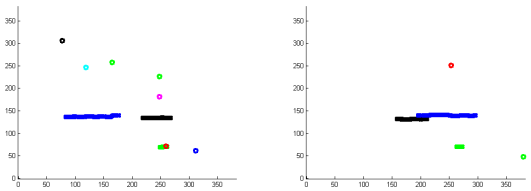


Figure 3. Two snapshots of running the proposed algorithm to track three targets in a simulation setting. Three tracked targets are highlighted with the history trajectories, and circle dots are clutters.

Algorithm (fps)	M=3	M=5	M=7	M=9
Ours	33.6	26.8	15.3	6.4
JPDA	28.3	9.5	2.1	0.35

Table 1. Average running time of the proposed approach v.s. JPDA.

We implemented the JPDA filter with the same Kalman Filter setting. By increasing the number of ground truth targets to track, the average running time of both algorithms are recorded for the comparison of computational complexity. Table 1 shows the average frame rates of two algorithms on tracking 3, 5, 7, 9 targets respectively. As can be observed from the table, the reduced computational cost gained by the proposed algorithm over JPDA becomes more and more obvious with the number of targets increased, which clearly demonstrates the scalability of the proposed algorithm to handle large number of targets.

4.2. Part-based Person Tracking

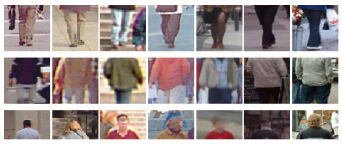


Figure 4. Positive samples of part-based human dataset. Top: legs; Middle: torso; Bottom: head-shoulder.

The proposed approach is also applied to tracking walking people in challenging real-world surveillance videos (CAVIAR) [1], where ground truth data is available enabling the quantitative performance evaluation. A part-based person representation, as discussed in Section 3.1, is used, where the part detectors are trained based on our adaptations of the well known SVM classifier with the Histogram of Oriented Gradient features [4]. Some representative positive samples are shown in Figure 4. Directly applying the learned classifiers produces many false alarms. Since the $3D$ site geometry is available in CAVIAR data, human height constraint is leveraged to reduce the classifiers' false alarm rates.

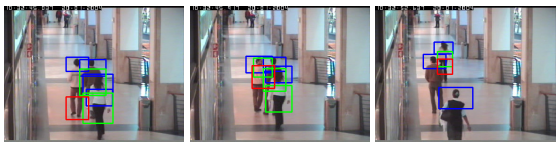
The tracker we applied is a $3D$ tracker. In our implementation, motion state of each tracker contains person's ground plane location and height information. The $3D$ knowledge also enables the system to use the depth-based data association priori in Section 3.2. The likelihood computation of each $3D$ tracker in Eq. 14 involves three steps, i.e., the projection of tracker's $3D$ bounding box to $2D$ image, collecting the corresponding parts' color histograms from the projected $2D$ bounding box, and then matching against the target part-based histogram models acquired and maintained during the tracker's lifespan. Due to such high nonlinearity, Kalman Filter becomes invalid, thus we adopt Particle Filter to run the tracker, i.e., the variational probability $Q_i(x_{i,t})$ is represented by a weighted particle set, and all the integral computations in previous sections become summations. Figure 5 shows our system results on one of the representative CAVIAR videos. The top row illustrates sample images of part-based detections after $3D$ geometric screening. The corresponding tracking results are shown at the bottom row of Figure 5. A unique ID is assigned to each tracker. It is clear that most of the people in these images have been successfully tracked, while only one of them, pointed out at the bottom-right image of Figure 5, is miss-tracked primarily due to a long period of occlusion, and then gets re-initialized with a new tracker ID after occlusion is complete. Figure 6 shows another interesting results. Thanks to the part-based target representation, a newly established tracker quickly picks up a person, pointed out in the figure, when the person is only partially presenting in the scene.

(a) Detection results by running the part-based SVM-HoG classifiers. Blue: head-shoulder detections; Green: torso detections; Red: leg detections.

(b) Tracking results of the proposed method with the depth-based association priori.

Figure 5. Detections and tracking in a crowded environment. Sequence 1: see the supplemental video for details.

To measure the system performance, quantitative studies are conducted. We measures the track-level performance, i.e., answering the question of how well we track targets. To measure this performance, we need to find the assignments between system generated tracks and ground truth tracks [1]. We solve this complex track-level assignment



(a) Detection results by running the part-based SVM-HoG classifiers. Blue: head-shoulder detections; Green: torso detections; Red: leg detections.

(b) Tracking results of the proposed method with the depth-based association priori.

Figure 6. Detections and tracking in a crowded environment. Sequence 2: see the supplemental video for details.

problems using greedy method. Once the assignments are returned, two quantitative measures proposed in [11] are computed, 1) track completeness factor (TCF), which measures on average what temporal ratio of a ground truth track is covered by system generated tracks. An ideal TCF score is 100%, i.e., successfully track a target during its complete lifespan. 2) track fragmentation factor (TFF), which tells on average how many system tracks are used to match one ground truth track. This factor implicitly correlates with the system's performance on keeping target identity during tracking. An ideal TFF score is 1. Table 2 reports the system performance using TCF and TFF on five selected CAVIAR videos. As shown by the TFF scores in the Table, our system reliably maintains target identity, considering the extremely challenging scenarios in CAVIAR data (See the Supplemental Video Submission for Details).

Seq ID	#1	#2	#3	#4	#5
# of Frames	1604	1520	1649	1376	1589
TCF	77.7%	81.2%	71.7%	80.3%	75.8%
TFF	1.35	1.18	1.67	1.23	1.33

Table 2. The track-level performance measure of the proposed approach. See text for details

In terms of computational efficiency, as supported by our theoretical studies in Section 2.4 and Section 3, the complexity of the proposed tracking algorithm is approximately quadratic to the number of targets being tracked, i.e., $O(M^2)$, due to the use of part-based representation in Section 3.1 and depth-based association priori in Section 3.2, which still saves tremendous computational cost compared to the joint state space approaches. More specifically, without counting the computations of running the part-based detections for all frames, the average running speed of the tracking algorithm itself is around 8fps for a moderately crowded scene on a Pentium 4-M 3GHz machine.

5. Conclusions

A novel distributed framework for multi-target tracking is proposed, where both trackers' motion and association variables take a decentralized representation. Formulate the multi-target tracking as a missing data problem, we analytically show that the posteriori distributions of trackers' motions and the optimal solution of trackers' data associations can both be iteratively computed in a variational EM framework. The tremendous generalization flexibility of the proposed method has also been shown.

References

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[2] Y. Bar-Shalom and T. Fortmann. *Tracking and Data Association*. Academic Press, Orlando, FL, 1988.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume I, pages 886–893, San Diego, CA, June 2005.

[5] C. Hue, J. Cadre, and P. Perez. Tracking multiple targets with particle filtering. *IEEE Trans on Aerospace and Electronic Systems*, 38(3):791–812, 2002.

[6] M. Isard and J. MacCormick. BraMBLe: A bayesian multiple-blob tracker. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 34–41, Vancouver, Canada, 2001.

[7] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 2000.

[8] Z. Khan, T. Balch, and F. Dellaert. An mcmc-based particle filter for tracking multiple interacting targets. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2004.

[9] O. Lanz. Approximate bayesian multibody tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 28(9):1436–1449, 2006.

[10] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple targets. In *In Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 572–578, Greece, 1999.

[11] A. G. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu. Multi-object tracking through simultaenous long occlusions and split-merge conditions. In *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, New York City, June 2006.

[12] C. Rasmussen and G. Hager. Probabilistic data association methods for tracking complex visual targets. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 560–576, 2001.

[13] J. Rittscher, P. H. Tu, and N. Krahnstoever. Simultaneous estimation of segmentation and shape. In *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, June 2005.

[14] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers. People tracking with a mobile robot using sample-based joint probabilistic data association filters. *International Journal of Robotics Research (IJRR)*, 22(2), 2003.

[15] H. Tao, H. S. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 24(1):75–89, 2002.

[16] Y. Weiss and W. T. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Trans on Information Theory*, 47(2), 2001.

[17] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *In Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, New York City, June 2006.

[18] T. Yu and Y. Wu. Collaborative tracking of multiple targets. In *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, Washington, D.C., June 2004.

[19] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, Washington, D.C., June 2004.