Real-time Tracking of Multiple Objects by Linear Motion and Repulsive Motion

Lejun Shen[1], Zhisheng You[2] and Qing Liu[1]

Chengdu Sport University[1], Sichuan University[2], China

Abstract. Successful multi-object tracking requires consistently maintaining object identities and real-time performance. This task becomes more challenging when objects are indistinguishable from one another. This paper presents a Bayesian framework for maintaining the identities of multiple objects. Our semi-independent joint motion model (SIMM) solves the coalescence and identity switching problem in real time. This joint motion model is a non-parametric mixture model that simultaneously captures linear motion and repulsive motion. Linear motion is a constant velocity model, while repulsive motion is described by a repulsive potential in MRF. By maintaining multimodality from multiple motion models, we can infer the appropriate motion model using image evidence and consequently avoid many identity switching errors. Moreover, we develop a new sampling method that does not suffer from the curse of dimensionality because of the availability of high-quality samples. Experimental results show that our approach can track numerous objects in real time and maintain identities under difficult situations.

1 Introduction

Multi-object tracking is important for many applications, such as video surveillance, robotics, radar-based tracking of aircraft, and sports video analysis. It is a relatively easy task when objects are distinguished from one another. In practical tracking applications, however, some objects are indistinguishable, such as similar looking people in surveillance videos [29], players of the same team in broadcast sport videos [15], and unlabelled measurements in radar tracking [26]. This paper aims to maintain the identities of multiple indistinguishable objects in real time.

The ambiguity caused by indistinguishable objects is one of the main difficulties encountered in multi-object tracking [6]. When objects present nonlinear motion, tracking becomes even more difficult, involving two subproblems: the *coalescence problem* (trackers associate more than one trajectory with some objects while losing track of others) and the *identity switching problem* (two intersecting trajectories exchange identities). The coalescence problem can be solved by exclusion constraints: (C1) detection response (measurement) should be assigned to at most one trajectory, or (C2) two objects cannot occupy the same space. Constraint (C1) is known as *data association* [17, 27, 16]. Constraint (C2) is known as *Markov random field (MRF) motion* [10, 30] in image plane or physical exclusion [13] in 3D world space. Once the coalescence problem is resolved, the identity switching problem becomes crucial, as discussed in the following paragraphs.

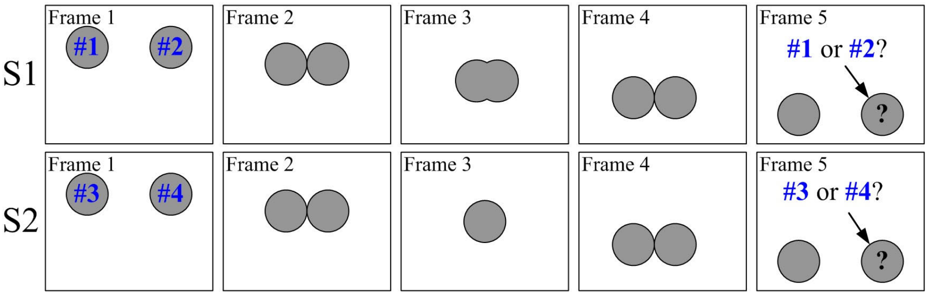

Fig. 1. We assign two identities to the two indistinguishable objects in frame 1. The correct identity of the object indicated by question mark depends on the motion model we *selected*, as shown in Fig.2.

A linear motion model is commonly used in multi-object tracking, i.e., a constant velocity model or motion affinity using the assumption of linear motion. As shown in [20, 29, 6], however, the linear motion model is often plagued by nonlinear motion patterns in real scenes. This problem can be shown by a simple example: sequence S1 in Fig.1, where two indistinguishable objects are marked ♯1 and ♯2 in frame 1. After frames 2, 3, and 4, what is the correct identity of the object indicated by the arrow in frame 5 ? Linear motion indicates that the identity is ♯1 (see H1 in Fig.2), but the image evidence points to the identity as ♯2 because full occlusion does not occur in frame 3. The absence of full occlusion implies that no intersection happens, making the nonlinear motion model the preferred tool for disambiguating the object identities shown in S1.

The nonlinear motion model is also used in multi-object tracking. The MRF motion model is a typical representative and is often used to describe the repulsive force of objects (hereinafter called *repulsive motion*). It can better explain the directional changes in S1, but repulsive motion also presents the identity switching problem. Consider sequence S2 in Fig.1. S2 also has 5 consecutive frames showing the continuous motions of two objects marked ♯3 and ♯4 in frame 1. After frames 2, 3, and 4, what is the correct identity of the object indicated by the arrow in frame 5? The repulsive motion model indicates that the correct identity is ♯4. The full occlusion in frame 3 may mean that intersection (H3 in Fig.2) and departure (H4 in Fig.2) are equally possible, but H3 can better explain the sequence in accordance with the inertia law of physics. Therefore, the linear motion model is the preferred tool for maintaining the identities of objects presented in S2.

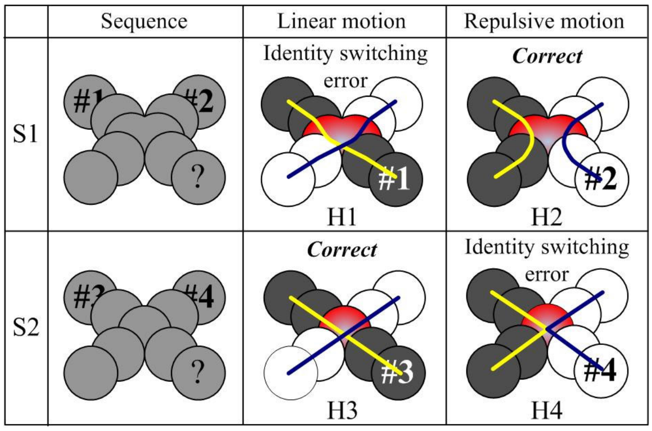

Fig. 2. Tracking results of linear motion and repulsive motion models. Our method can naturally evolve a correct motion model from multiple models.

How to select the correct motion model (i.e. repulsive motion model in S1 and linear motion model in S2)? We use a Bayesian framework that adaptively selects an appropriate motion model. We generate multiple hypotheses according to multiple motion models, evaluate the likelihood of these hypotheses giving rise to observed data, maintain the multiple hypotheses, and infer the correct identity using the motion model that is consistently supported by the observed data.

The first contribution of this paper is a semi-independent joint motion model (Section 3.1). It is a non-parametric mixture model, which takes into account the repulsive force between objects, as well as their inertial force. This mixed motion model removes the ambiguity caused by indistinguishable objects. Moreover, it maintains multimodality (or multiple hypotheses), thereby evolving an correct motion model and avoiding identity switching errors (Section 3.2).

The second contribution of this paper is an efficient sampling method that does not suffer from the curse of dimensionality. The advantage of Monte Carlo simulation is that the accuracy of the Monte Carlo estimator does not depend on the dimensionality of a problem. High accuracy may be achieved with a relatively small number of high-quality samples. We design a tractable importance distribution to generate high-quality samples (Section 3.3). Therefore, our method scales well even under difficult situations (e.g. severe occlusion).

2 Related Works

Multi-object tracking methods can be divided into two categories. The first uses only current and past information to estimate the current state [22, 17, 25, 19, 10, 30, 26, 12, 23, 4, 3]. It is well suited for real-time applications, but cannot recover from failure, because data association decisions are based only on past information. These decisions, once made, are fixed and may later be revealed as suboptimal. Maintaining multiple hypotheses can delay data association decision making until enough information has been obtained to derive the optimal solution. The capability of maintaining multiple distributions (p_L and p_R in this paper) is the key difference between our method and others. We use a non-parametric mixture model [25] to facilitating identity maintenance where the standard particle filter fails. Besides, we propose a novel importance distribution to avoid the curse of dimensionality.

The second category of multi-object tracking also uses future information to estimate the current state within a given time window [21, 27, 13, 28, 29, 2, 15, 6, 16]. It more effectively overcomes the ambiguities caused by long-time occlusions and false or missed detections. Initialization and termination are fully automatic [24]. The accuracy of object detectors, however, remains far from perfect. Detectors may generate unreliable detection responses if an object is fully or partially occluded by others. These detection errors propagate to the tracking (or data association) module and consequently cause identity switching error. For example, the comparison of S1 and S2 shows that the key difference is frame 3 (highlighted in red in Fig.2). The missed or inaccurate detection response at frame 3 damages this key evidence and introduce identity ambiguity. Tracking-by-detection approaches suffers from this ambiguity. To avoid this problem, we use image patch, instead of detection response in the observation model, because image-patch based observation model capture the difference between S1 and S2.

The nonlinear motion model has been used in multi-object tracking applications, such as social behavior of people (path planning [20], moving groups [3], game context features [15]) and knowledge about scenes (motion patterns [29], entry/exit points [27]). By contrast, MRF motion model does not require the knowledge of scenes or targets. Khan et al. [10] modeled the interaction between objects by MRF motion model, and Yu et al. [30] proposed a set of collaborative trackers to solve the coalescence problem. Lanz [12] proposed a hybrid joint-separable filter, which is similar to our approach. Qu et al. [23] proposed a distributed architecture. Khan et al. demonstrated the failure mode of MRF motion when basic assumption (C2) is violated. High-order motion models can avoid this failure [23]. In section 3.2, we discuss why (C2) causes failure and how to improve it using our SIMM approach.

Data association is naturally a discrete problem and trajectory estimation is a continuous problem [2]. The inference in discrete-continuous (hybrid) model is NP-hard [14]. The MRF motion model is described in a continuous state space and its repulsive potential is also a continuous function [10, 30]. This continuous nature makes designing a tractable importance distribution possible. Our importance distribution Q^{simm} drastically reduces computational effort (Section 3.3).

Khan et al. [9] use multiple nearly independent trackers (a real-time version of [10]). Hess et al. [7] devised a pseudo-independent log-linear filter. They used previous states $\hat{x}_{t-1}$ of other objects to compute interaction features between objects. This simplification requires less computation but leads to identity switching error when severe occlusion presents because of the sample impoverishment. Our importance distribution Q^{simm} generate high-quality samples and consequently avoid sample impoverishment.

The mixture of multiple motion models is not a new idea. Isard et al. used a manual transition matrix in mixed-state CONDENSATION [8]. Yu et al. used a binary performance indicator to switch between motion models [31]. Oh et al. used a semi-Markov transition matrix learned from data [18]. Kwon et al. decomposed the motion of an object into two kinds of motions [11]. We do not manually set a transition matrix [8], nor empirically set a threshold of performance indicator [31], nor learn model parameters from data [18, 7, 20, 15, 6]. Our model parameters (m_0 and m_1) are fixed all the time. The particle sets can naturally evolve a correct motion model from multiple motion models according to image evidence.

3 Semi-Independent Multi-object Tracking

3.1 Problem Formulation

Multi-object tracking can be formalized as a sequential Bayesian estimation problem. We denote the state of the i^{th} object at time t by $x_{i,t}$, the joint state by $X_t = \{x_{1,t}, x_{2,t}, ..., x_{M,t}\}$, the local observation of $x_{i,t}$ by $y_{i,t}$, and the joint observation by $Y_t = \{y_{1,t}, y_{2,t}, ..., y_{M,t}\}$, with M number of objects.

For computational efficiency, we assume that joint observation model $p(Y_t|X_t)$ can be factorized with respect to each individual object.

$$p(Y_t|X_t) = \prod_i p(y_{i,t}|x_{i,t}) \tag{1}$$

To maintain multiple distributions of multiple motion models, we formulate joint motion model $p(X_t|X_{t-1})$ as a mixture model, i.e.,

$$p(X_t|X_{t-1}) = \left(m_0 + m_1 c_G \prod_{j>i} \varphi(x_{i,t}, x_{j,t}) \right) \prod_i p(x_{i,t}|x_{i,t-1}) \tag{2}$$

where m_0 and m_1 are the mixture weights with m_0+m_1=1. c_G is a normalization factor. $p(x_{i,t}|x_{i,t-1})$ is a constant velocity model. $\varphi(x_{i,t}, x_{j,t})$ denotes the repulsive potential that solves the coalescence problem. With the use of different parameter values, the motions of multiple objects may be completely independent (m_0=1,m_1=0), dependent (m_0=0,m_1=1) [30, 10], or *semi-independent* (m_0=m_1=0.5 in this paper). The intuition behind this model is that a target $x_{i,t}$ behaves in accordance not only with its own past state $x_{i,t-1}$ with weight m_0, but also with the behaviors of other targets $x_{j,t}$ with weight m_1.

This semi-independent joint motion model is the key novelty of this paper. The component $(m_0 + m_1 c_G \prod_{j>i} \varphi(x_{i,t}, x_{j,t}))$ is designed to disambiguating object identities (Section 3.2). The component $\prod_i p(x_{i,t}|x_{i,t-1})$ outside the bracket is the key to real-time performance and scalability (Section 3.3). After using the assumption [12] that the joint distributions are approximated via the outer product of their marginal components $p(X_{t-1}|Y_{t-1}) = \prod_i p(x_{i,t-1}|Y_{t-1})$, equations (1) and (2) lead to an multi-object tracking framework as follows.

Prediction:

$$p(x_{i,t}|Y_{t-1}) = \int p(x_{i,t}|x_{i,t-1})p(x_{i,t-1}|Y_{t-1})dx_{i,t-1} \tag{3}$$

where $p(x_{i,t}|x_{i,t-1})$ is a is a constant velocity model.

Updating:

$$p_L(x_{i,t}|y_{i,t}) = c_i p(y_{i,t}|x_{i,t})p(x_{i,t}|Y_{t-1}) \tag{4}$$

where c_i represents normalization factors and p_L denotes the posterior of linear motion.

Joint updating:

$$p_{MRF}(X_t|Y_t) = c_G \prod_{j>i} \varphi(x_{i,t}, x_{j,t}) \prod_i p_L(x_{i,t}|y_{i,t}) \tag{5}$$

where p_{MRF} denotes the joint distribution of MRF.

Marginalization:

$$p_R(x_{i,t}|Y_t) = \int p_{MRF}(X_t|Y_t)dx_{\neg i,t} \tag{6}$$

where $x_{\neg i,t}$ represents vector X_t with the i^{th} component removed and p_R denotes the posterior of repulsive motion.

Mixture of distributions:

$$p(x_{i,t}|Y_t) = m_0 p_L(x_{i,t}|y_{i,t}) + m_1 p_R(x_{i,t}|Y_t) \tag{7}$$

Please see our supplementary material (*.pdf) for complete mathematical derivations.

3.2 Mixture Model

We discuss how posterior distributions p_L and p_R collaboratively maintain object identities.

Likelihood ambiguity occurs if multiple interacting objects are identical in appearance (e.g., the tennis balls in Fig.3). The observation model is useless in inferring the object identities. This ambiguity causes multiple modes in distribution $p_L(x_2)$ at t=70 and consequently leads to the coalescence problem. By contrast, $p_R(x_2)$ has a single peak from t=56 to t=80 even when $p_L(x_2)$ has multiple modes. Therefore, *repulsive potential* $\varphi(\cdot)$ *eliminates the likelihood ambiguity that originates from the observation model.* As a result, x_2 maintains the correct identity throughout the interaction. Note that *no* complete occlusion occurs in S1 and the nearest distance between two true trajectories is 4 pixels (overlap ratio 74.8%) at t=67.

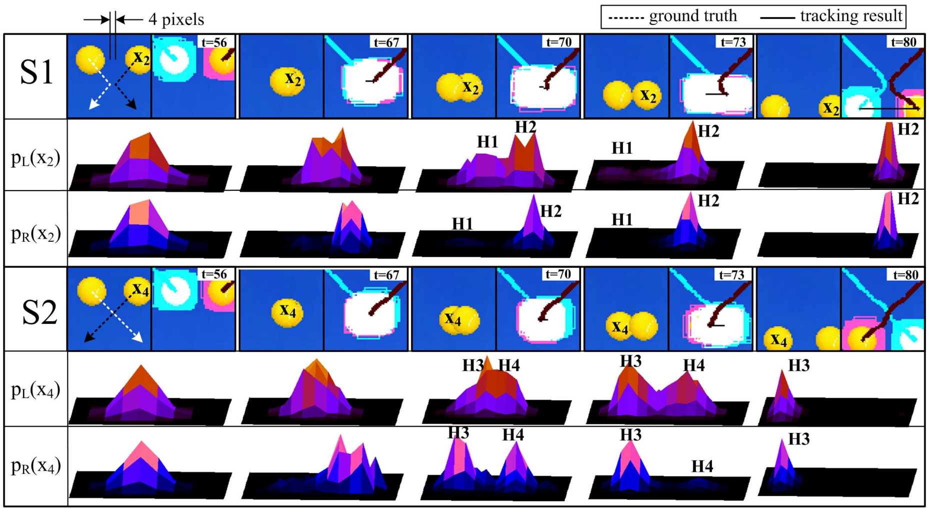

Fig. 3. The source image and tracking result of S1 are shown in the 1st row. The posterior of linear motion p_L and repulsive motion p_R are shown in the 2nd and 3rd row. Please see the text and our supplementary material for more details.

Identity ambiguity occurs if assumption (C2) is violated [10]. *Complete occlusion leads to the identity ambiguity* in S2, subsequently causing the multimodality of $p_R(x_4)$ at t=70 (Fig.3). In other words, two hypotheses exist: x_4 is associated with left-hand side peak H3 and right-hand side peak H4. This multimodality is maintained from t=67 to t=78 by mixture model (2). Finally, H3 is consistently supported by the observations in this time interval. Meanwhile, H4 is pruned away by $\varphi(\cdot)$. As a result, x_4 derives the correct identity at t=78. By contrast, the traditional MRF motion model votes only H4, resulting in identity switching error. In this situation, *linear motion helps solve the identity ambiguity that stems from MRF motion.*

In summary, both linear motion and repulsive motion are indispensable to identity maintenance. Maintaining multimodality delays data association decision making until enough observations are collected.

3.3 Monte Carlo Implementation

We present a novel Monte Carlo implementation of the tracking framework in Section 3.1.

Curse of dimensionality. Computer vision systems are confronted with demands to track numerous objects in dense scenarios. In equation (5), the dimension of joint state X_t is high. Exploring a high-dimensional joint state space is generally computationally intensive. Fortunately, high-dimensional problems can be efficiently solved by Monte Carlo simulation if high-quality samples are available. The factorization property of equation (1) and the continuous nature of $\varphi(\cdot)$ enable the convenient use of importance sampling. In particle filter literature, the optimal importance distribution is $Q^{opt}(X_t|Y_t, X_{t-1})$. Transition prior (or joint proposal distribution in [10])

$$Q^{transition-prior} = \prod_i \int p(x_{i,t}|x_{i,t-1})p(x_{i,t-1}|Y_{t-1})dx_{i,t-1} \tag{8}$$

is inefficient. We propose a novel importance distribution

$$Q^{simm} = \prod_i p_L = \prod_i p(y_{i,t}|x_{i,t}) \int p(x_{i,t}|x_{i,t-1})p(x_{i,t-1}|Y_{t-1})dx_{i,t-1} \tag{9}$$

to approximate optimal importance distribution Q^{opt}. Q^{simm} provides us high-quality joint state samples because it incorporates current observations $y_{i,t}$.

Suppose that $p(x_{i,t-1}|Y_{t-1})$ at the previous time step is approximated by a set of N weighted particles:

$$p(x_{i,t-1}|Y_{t-1}) \approx \left\{x_{i,t-1}^n, \pi_{i,t-1}^n\right\}_{n=1}^N \tag{10}$$

where N is the number of particles, n is the index of samples, and $\pi_{i,t-1}^n$ is the weight of the n^{th} particle.

Prediction. The prediction is generated by a proposal density that incorporates current detection responses D_t and transition prior $p(x_{i,t}|x_{i,t-1})$ [19].

$$q_i = (1-\lambda)p(x_{i,t}|x_{i,t-1}) + \lambda g(x_{i,t}, D_t) \tag{11}$$

where $p(x_{i,t}|x_{i,t-1})$ is a constant velocity model, $g(x_{i,t}, D_t) \sim N(x_{i,t} - D_t; \sigma^2)$ denotes the normal distribution evaluated for the distance between D_t and $x_{i,t}$. q_i improves tracker robustness and reduces model drift. A lower value of λ implies a suppression of false positive detection.

Updating. We derive the posterior of linear motion according to observation model $p(y_{i,t}|x_{i,t})$:

$$p_L(x_{i,t}|y_{i,t}) \approx \left\{x_{i,t}^n, L_{i,t}^n = p(y_{i,t}|x_{i,t}^n)\right\}_{n=1}^N \tag{12}$$

where $L_{i,t}^n$ is the linear motion weight of the n^{th} particle.

Joint updating. We draw K un-weighted samples

$$p_L(x_{i,t}|y_{i,t}) \approx \left\{x_{i,t}^{k*}, B_k(i)\right\}_{k=1}^K \tag{13}$$

by resampling from equation (12) for $i = 1, ..., M$. $B_k(i)$ refers to the index of the particle in (12) at the k^{th} sampling [22]. Storing the index of its parent is unnecessary in the standard particle filter, but it is useful in later procedures. We approximate Q^{simm} by combining M independent un-weighted samples to one joint state sample:

$$Q^{simm} \approx \left\{X_t^{k*}, B_k\right\}_{k=1}^K \tag{14}$$

where $X_t^{k*} = \left(x_{1,t}^{k*}, ..., x_{M,t}^{k*}\right)$, $B_k = (B_k(1), ..., B_k(M))$, and K is the number of joint state samples. Then, $p(X_t|Y_t)$ is approximated by

$$p_{MRF}(X_t|Y_t) \approx \left\{ X_t^{k*}, B_k, w_t^k = \prod_{j>i} \varphi(x_{i,t}^{k*}, x_{j,t}^{k*}) \right\}_{k=1}^{K} \tag{15}$$

where w_t^k is the weight of the k^{th} joint state sample and the repulsive potential function is defined as

$$\varphi(x_i, x_j) = \exp\left(-\alpha \times \left(\frac{overlap(x_i, x_j)^2}{area(x_i)area(x_j)}\right)^2\right) \tag{16}$$

where $overlap(x_i, x_j)$ is the number of pixels that overlap between x_i and x_j , and $area(x_i)$ is the number of pixels of x_i. Note that (16) is currently implemented in the image-plane, but it can be easily used in 2D-ground-plane, 3D-world-space or pose-space after replacing (16) with a new function.

In visual tracking, the most computationally expensive operation is likelihood evaluation (12). The computational cost of resampling (13) is much lower than likelihood evaluation (12). For example, the computational time is 0.15 ms per sample in equation (12) and 0.004 ms per sample in equation (13). To generate 4000 joint state samples, traditional importance sampling needs 600 ms (=4000*0.15). Our joint updating needs only 46 ms (=200*0.15+4000*0.004) if K=200. Moreover, the resampling preserves the support of distribution, and therefore avoids sample impoverishment.

Marginalization. An advantage of Monte Carlo simulation is that some computations are particularly easy. Marginalization is a good example. We can obtain marginal samples x^i by sampling (x^i, u^i) from augmented distribution $p(x, u)$, and disregard the u^i component [1]. In equation (6), $x_{\neg i,t}$ is an auxiliary variable. Given (15), $P_R(x_{i,t}|Y_t)$ can be easily obtained by ignoring $x_{\neg i,t}$

$$p_R(x_{i,t}|Y_t) \approx \left\{x_{i,t}^{k*}, B_k(i), w_t^k\right\}_{k=1}^{K} \tag{17}$$

Mixture of distributions. The posterior of linear motion (12) and repulsive motion (17) cannot be directly mixed because N, the number of particles, is not equal to K, the number of joint state samples. We accumulate and normalize the weights according to $B_k(i)$:

$$p_R(x_{i,t}|Y_t) \approx \left\{ x_{i,t}^n, R_{i,t}^n = \sum_{B_k(i)=n} w_t^k \Bigg/ \sum_{k=1}^{K} w_t^k \right\}_{n=1}^{N} \tag{18}$$

where $R_{i,t}^n$ is the repulsive motion weight of the n^{th} particle.

Equation (18) is a normalized histogram representation of equation (17). Once histogram (18) has been computed, dataset (17) can be discarded, which can be advantageous if $K >> N$. Second, the transfer from equations (17) to (18) does not change the expectation of Monte Carlo estimation. Third, computing mixture model (7) is easy.

$$p(x_{i,t}|Y_t) \approx \left\{x_{i,t}^n, \pi_{i,t}^n = m_0 \times L_{i,t}^n + m_1 \times R_{i,t}^n\right\}_{n=1}^N \tag{19}$$

Note that equations (9), (14), (15), (18), (19) are new ideas of this paper. Normalization (c_i and c_G) is very important in our algorithm, which help us to improve the inference of data association decision-making. We explicitly implemented them (line 19 and 20 in Algorithm 1).

Algorithm 1 semi-independent joint motion model (SIMM) particle filter

```
 1: Input: $\left\{x_{i,t-1}^n, \pi_{i,t-1}^n\right\}_{n=1}^N$
 2: for $i = 1$ to $M$ do
 3:    Resample $\{x_{i,t-1}^n\}_{n=1}^N$ from $\left\{x_{i,t-1}^n, \pi_{i,t-1}^n\right\}_{n=1}^N$
 4:    Sample $\{x_{i,t}^n\}_{n=1}^N$ from $q_i$, by Eq. (11)
 5:    $L_{i,t}^n = p(y_{i,t}|x_{i,t}^n),\ for\ n = 1, ..., N$, by Eq. (12)
 6:    $R_{i,t}^n = 0,\ for\ n = 1, ..., N$
 7: end for
 8: for $k = 1$ to $K$ do
 9:    $w = 1.0$
10:    Sample index $B(i)$ from the discrete distribution given by $\left\{x_{i,t}^n, L_{i,t}^n\right\}_{n=1}^N,\ for\ i =$
       $1, ..., M$, by Eq. (9) or (14)
11:    for $i = 1$ to $M$ do
12:       for $j = i + 1$ to $M$ do
13:          $w = w \times \varphi(x_{i,t}^{B(i)}, x_{j,t}^{B(j)})$, by Eq. (15)
14:       end for
15:    end for
16:    $R_{i,t}^{B(i)} = R_{i,t}^{B(i)} + w,\ for\ i = 1, ..., M$, by Eq. (18)
17: end for
18: for $i = 1$ to $M$ do
19:    Normalize weights $L_{i,t}^n,\ for\ n = 1, ..., N$
20:    Normalize weights $R_{i,t}^n,\ for\ n = 1, ..., N$
21:    $\pi_{i,t}^n = m_0 \times L_{i,t}^n + m_1 \times R_{i,t}^n\ ,\ for\ n = 1, ..., N$, by Eq. (19)
22: end for
```

4 Experiments

We evaluate our approach on six sequences: four synthetic video sequences (S1, S2, S3, and S4) and two public video sequences (PETS2009 S2L1 and UBC Hockey). We compare our method with one data association method (MC-JPDAF [26]) and four MRF motion methods: MCMC-based particle filter (MCMC) [10], Mean Field Monte Carlo (MFMC) [30], Distributed Multiple Object Tracking (DMOT) [23] and Psuedo-Independent Log-Linear Filters (PILLFs) [7]).

Experimental setup. We use 6000 particles, propagate 10 samples to the next time step, and discard 25% to burn-in in MCMC; N=400 and 5 iterations are run in MFMC; N=400 and 6 iterations are run in DMOT. We use K=4000, N=400 and λ=0.05 in our method.

The observation model of MC-JPDAF is the distance between the particle and the associated detection using Gaussian distribution. MCMC, MFMC, D-MOT, and our method use the same observation model, where $p(y_{i,t}|x_{i,t})$ is the Bhattacharyya distance between HSV color histograms, and the same potential (16) with α=16.

In PILLFs, we use a constant velocity model and use two feature functions. The appearance feature f_1 is the Bhattacharyya distance between HSV color histograms and the interaction feature f_2 is (16). For fair comparisons, we do not use the error-driven discriminative training in [7]. Hence, the feature weights are fixed and equal $w_1 = w_2 = 0.5$.

In the PETS2009 dataset, we use a publicly available pedestrian detector [5]. In the UBC Hockey sequence, the detector is taken from [19]. In the synthetic sequences, the tracker is manually initialized in the first frame and an image background subtraction is used to detect BLOB-like objects.

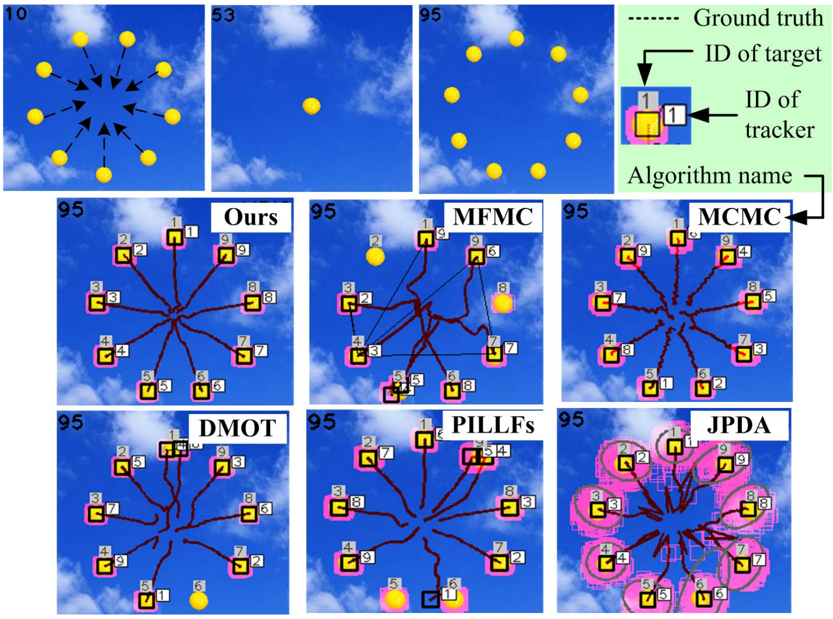

Fig. 4. Source images and tracking results for S3 (severe occlusion). Refer to text and supplementary material for more details.

4.1 Identity Switching Error

Identity switching is one of many important metrics in multi-object tracking [27]. This metric quantitatively evaluates the intrinsic ability of a tracking algorithm. It can be easily recognized if the "ID of target" does not equal the "ID of tracker" in Fig.4.

Severe occlusion (S3). The most serious occlusion of 9 objects is that they merge into a single measurement (Fig.4). This sequence challenges many existing methods, because 8 objects are fully occluded at t=53. MC-JPDAF has no identity switching error bacause objects move with the linear motions. MCMC, MFMC, and DMOT fail because their basic assumption (C2) is violated. PILLFs fail because it uses previous states $\hat{x}_{t-1}$ to compute interaction features, not all the particles.

As shown in Table 1, MCMC, MFMC, DMOT and PILLFs all effectively work for S1, which fits their basic assumption (C2). MC-JPDAF effectively works for S2 and S3, which fit the linear motion assumption that underlies this algorithm. Table 1 shows that our tracking framework produces better results than Linear model (MC-JPDAF), MRF model (MFMC, MCMC, DMOT) and Log-Linear model (PILLFs), because our motion model (2) is less dependent on a single assumption than are the other models.

VAR1 is a variant of our method with λ=0.2. The high value of λ makes it suffers from the unreliable detection responses. VAR2 is a variant of our method with m_0=0, which blocks the contribution of linear motion model and suffers from identity switching problem in S2.

S2L2 sequence from PETS2009. It has 795 frames with a resolution of 768×576 and contains various occlusions, such as a long-time partial occlusion (t=36 to 368, ID=4,5), abrupt changes in direction (or repulsive motion) (e.g., t=21 to 44, ID=3), linear motion (e.g., t=77 to 134, ID=3), full occlusion between objects, and occlusion with a road sign. We use λ=0.05 to avoid model drift. The higher λ (VAR1) works well because objects are distinguishable.

In the soccer game sequence (Fig.5), severe occlusion presents from t=233 to 300 between two identical players. Our method has identity switching errors, which will be discussed in the section 5.

4.2 Computation Speed

All the experiments are implemented in C++ and run on an Intel Core2 Duo 1.66 GHz PC.

Severe occlusion (S3). In Table 2, R is the quotient of Max. divided by Min. The large value of R indicates poor scalability. The minimal computation time of MC-JPDAF occurs at frame 10 with 512 data association hypotheses. The maximal computation time occurs at frame 54 with 17,572,114 hypotheses, because the gating procedure is disabled. Thus, MC-JPDAF suffers from the curse of dimensionality. Moreover, the computational time of MFMC is high, because MFMC uses message passing and the MRF is a fully connected graph at frame 53 in S3. DMOT and FILLPs scales well.

Method	**Sequences**			
	S1	S2	S3	PETS2009 S2L1
MFMC [30]	0	1	>1	>1
MCMC [10]	0	1	>1	>1
MC-JPDAF [26]	1	0	0	>1
DMOT [23]	0	1	>1	>1
PILLFs [7]	0	1	>1	>1
Andriyenko et al. [2]	-	-	-	10
Yang et al. [29]	-	-	-	0
Segal et al. [24]	-	-	-	4
VAR1 (λ=0.2)	1	0	>1	0
VAR2 (m_0=0)	0	1	>1	>1
Ours (λ=0.05, m_0=0.5)	0	0	0	0

Table 1. Counting of identity switching errors on 4 sequences.

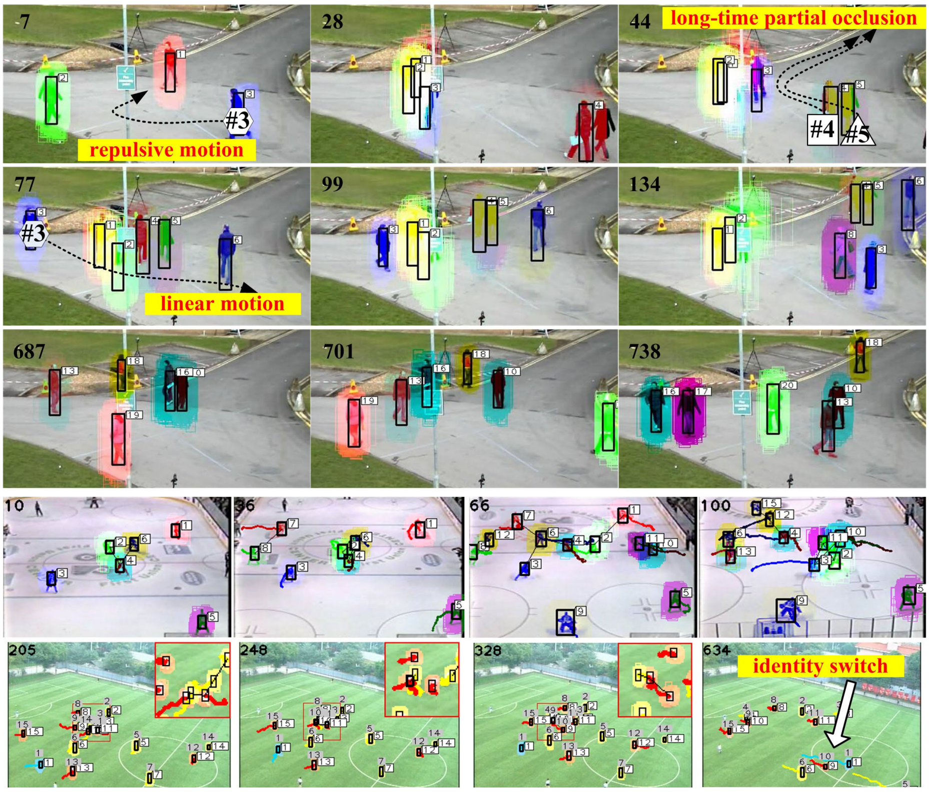

Fig. 5. Tracking results for PETS2009 S2L1 (1st, 2nd and 3rd rows), hockey [19] (4th row), and soccer game (5th row) sequences. Identity switching error is indicated by arrows. Refer to text and supplementary material for more details.

Method	Min.	Max.	Avg.	R
MFMC [30]	5.6	7910.5	2506.7	1412.5
MCMC [10]	49.5	62.8	53.1	1.2
MC-JPDAF [26]	1.7	2878.6	168.8	1693.3
DMOT [23]	5.6	25.3	9.5	4.5
PILLFs [7]	6.8	15.2	7.8	2.2
Ours	7.0	22.1	10.5	3.2

Table 2. Minimal, maximal, and average computation time (millisecond per frame) in S3 with R=Max./Min.

Long-time random sequence (S4). A total of M=36 objects (20×20 pixels) randomly walk in this video. It is a challenge because the dimension of the joint state is very high. Many complex interactions occur, including nonlinear motion before/within/after occlusion, and long-time occlusion. Our method (N=300, K=4000) can track 36 objects at 8.6 fps.

In our method, K is an empirically small constant and fixed all the time. Thus, our method scales well as the state space dimension M increases. Given the detector output, the speed of our method is 12 fps for PETS2009 S2L1 and 18.2 fps for UBC Hockey. Our executable program can be found in the supplementary material.

5 Conclusion

We have described and demonstrated a scalable real-time tracking framework for maintaining the identities of multiple indistinguishable objects. The motion model that is consistently supported by the observed data facilitates the inference of the correct object identities. Repulsive motion eliminates the ambiguity from observation models, but exhibits identity switching problems when complete occlusion occurs. Linear motion helps solve this problem. We use a mixture model to simultaneously capture these two motions and maintain multimodality. The proposed joint motion model is less dependent on a single assumption than are the other methods. Moreover, the factorization property and continuous nature of MRF motion enables the design of a tractable importance distribution, which generates high-quality samples and ensures the scalability of the algorithm.

The limitation of our method is that the identity ambiguity, caused by both complete and long-term occlusion between indistinguishable objects, cannot be reliably resolved because neither the observation model nor the motion model can provide sufficient evidence for inferring identities (Fig.5).

Acknowledgement. This work was supported by National Natural Science Foundation of China (61001195) and Technology Research and Development Program of Sichuan Province of China (2010JY0078). We would like to thank anonymous reviewers of this paper for helpful comments.

References

1. Andrieu, C., De Freitas, N., Doucet, A., Jordan, M.: An introduction to MCMC for machine learning. Machine learning **50** (2003) 5–43
2. Andriyenko, A., Schindler, K., Roth, S.: The Discrete-continuous optimization for multi-target tracking. Computer Vision and Pattern Recognition (2012)
3. Bazzani, L., Cristani, M., Murino, V.: Decentralized particle filter for joint individual-group tracking. Computer Vision and Pattern Recognition (2012)
4. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multiperson tracking-by-detection from a single, uncalibrated camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011) 1820–1833
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition (2005)
6. Dicle, C., Camps, O.I., Sznaier, M.: The Way They Move: Tracking Multiple Targets with Similar Appearance. International Conference on Computer Vision (2013)
7. Hess, R., Fern, A.: Discriminatively trained particle filters for complex multi-object tracking. Computer Vision and Pattern Recognition (2009)
8. Isard, M., Blake, A.: A mixed-state condensation tracker with automatic model-switching. International Conference on Computer Vision (1998)
9. Khan, Z., Balch, T., Dellaert, F.: Efficient particle filter-based tracking of multiple interacting targets using an MRF-based motion model. International Conference on Intelligent Robots and Systems (2003)
10. Khan, Z., Balch, T., Dellaert, F.: MCMC-based particle filtering for tracking a variable number of interacting targets. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1805–1819
11. Kwon, J., Lee, K.M.: Visual tracking decomposition. Computer Vision and Pattern Recognition (2010)
12. Lanz, O.: Approximate bayesian multibody tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006) 1436–1449
13. Leibe, B., Schindler, K., Cornelis, N., Van Gool, L.: Coupled object detection and tracking from static cameras and moving vehicles. IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008) 1683–1698
14. Lerner, U., Parr, R.: Inference in hybrid networks: Theoretical limits and practical algorithms. Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence (2001)
15. Liu, J., Carr, P., Collins, R.T., Liu, Y.: Tracking Sports Players with Context-Conditioned Motion Models. Computer Vision and Pattern Recognition (2013)
16. Milan, A., Schindler, K., Roth, S.: Detection-and Trajectory-Level Exclusion in Multiple Object Tracking. Computer Vision and Pattern Recognition (2013)
17. MacCormick, J., Blake, A.: A probabilistic exclusion principle for tracking multiple objects. International Journal of Computer Vision **39** (2000) 57–71
18. Oh, S., Rehg, J., Balch, T., Dellaert, F.: Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. International Journal of Computer Vision **77** (2008) 103–124
19. Okuma, K., Taleghani, A., Freitas, N., Little, J., Lowe, D.: A boosted particle filter: Multitarget detection and tracking. Europeon Conference on Computer Vision (2004)
20. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. International Conference on Computer Vision (2009)

21. Perera, A., Srinivas, C., Hoogs, A., Brooksby, G., Hu, W.: Multi-object tracking through simultaneous long occlusions and split-merge conditions. Computer Vision and Pattern Recognition (2006)
22. Pitt, M., Shephard, N.: Filtering via Simulation: Auxiliary Particle Filters. International Journal of Computer Vision **94** (1999) 590–599
23. Qu, W., Schonfeld, D., Mohamed, M.: Real-time distributed multi-object tracking using multiple interactive trackers and a magnetic-inertia potential model. IEEE Transactions on Multimedia **9** (2007) 511–519
24. Segal, A.V., Reid, I.: Latent Data Association: Bayesian Model Selection for Multi-target Tracking. International Conference on Computer Vision (2013)
25. Vermaak, J., Doucet, A., Prez, P.: Maintaining multimodality through mixture tracking. International Conference on Computer Vision (2003)
26. Vermaak, J., Godsill, S., Perez, P.: Monte carlo filtering for multi target tracking and data association. IEEE Transactions on Aerospace and Electronic Systems **41** (2005) 309–332
27. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. International Journal of Computer Vision **75** (2007) 247–266
28. Xing, J., Ai, H., Lao, S.: Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. Computer Vision and Pattern Recognition (2009)
29. Yang, B., Nevatia, R.: Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. Computer Vision and Pattern Recognition (2012)
30. Yu, T., Wu, Y.: Collaborative tracking of multiple targets. Computer Vision and Pattern Recognition (2004)
31. Yu, T., Wu, Y.: Decentralized multiple target tracking using netted collaborative autonomous trackers. Computer Vision and Pattern Recognition (2005)