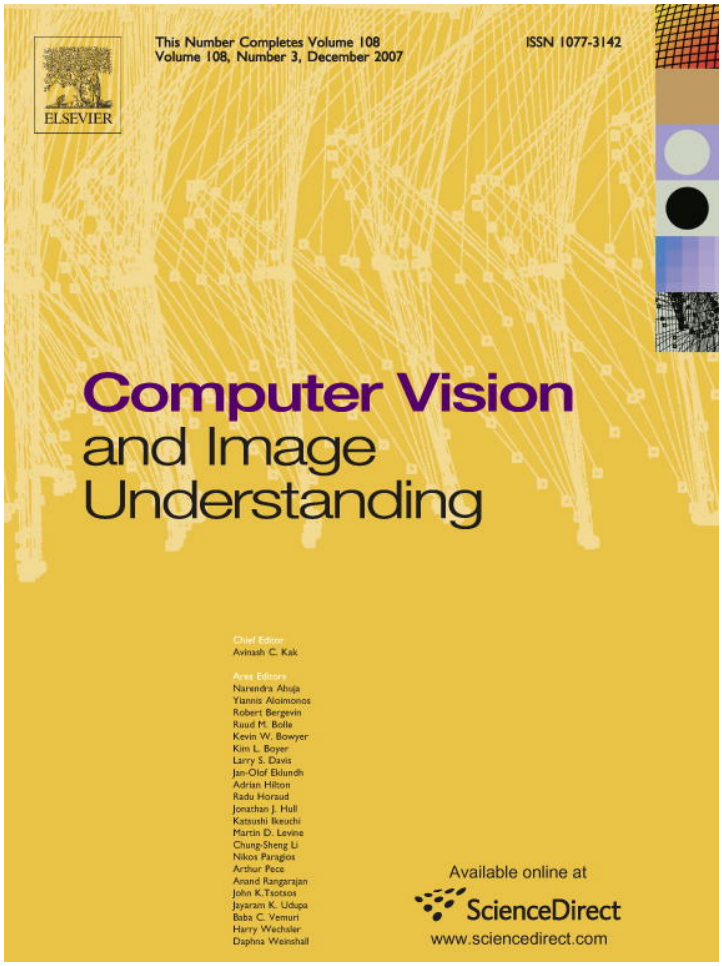

Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 108 (2007) 272–283
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

A decentralized probabilistic approach to articulated body tracking

Gang Hua [a,*], Ying Wu [b,1]

[a] Microsoft Live Labs Research, One Microsoft Way, Redmond, WA 98052, USA
[b] Department of Electrical Engineering and Computer Science, 2145 Sheridan Road, Northwestern University, Evanston, IL 60208, USA

Received 7 May 2005; accepted 5 November 2006
Available online 24 February 2007
Communicated by James MacLean

Abstract

We present a novel decentralized probabilistic approach to visual tracking of articulated objects. Analyzing articulated motion is challenging because (1) the high degrees of freedom potentially demands tremendous computation, and (2) the solution is confronted by the numerous local optima existed in a high dimensional parametric space. To ease these problems, we propose a decentralized approach that analyzes limbs locally and reinforces the spatial coherence among them at the same time. The computational model of the proposed approach is based on a dynamic Markov network, a generative model which characterizes the dynamics, the image observations of each individual limb, as well as the spatial coherence among them. Probabilistic mean field variational analysis provides an efficient computational diagram to obtain the approximate inference of the motion posteriors. We thus design the mean field Monte Carlo (MFMC) algorithm, where a set of low dimensional particle filters interact with one another and solve the high dimensional problem collaboratively. We also present a variational maximum a posteriori (MAP) algorithm, which has a rigorous theoretic foundation, to approach to the optimal MAP estimate of the articulated motion. Both algorithms achieve linear complexity w.r.t. the number of articulated subparts and have the potential of parallel computing. Experiments on human body tracking demonstrate the significance, effectiveness and efficiency of the proposed methods.

Keywords: Articulated body motion; Dynamic Markov network; Mean field Monte Carlo; Variational inference; Maximum a posteriori estimation

1. Introduction

Tracking articulated motion in video is an important problem, especially when the research of video-based human sensing has been advocated to achieve such emerging applications such as non-invasive perceptual human computer interfaces [1,2], intelligent video surveillance [3,4], gait analysis [5,6], automatic hand gesture recognition [7,8] and automatic video footage annotation [9], etc. The problem involves the localization and identification of a set of linked but articulated limbs. Inheriting all the difficulties from single object tracking, the problem of tracking articulated body has to tackle some special challenges. Some of them are the complications incurred by the high degrees of freedom of the articulated body: first, the computational complexity may increase exponentially with the increase of the dimensionality; and second, obtaining the optimal solution in a very high dimensional space is confronted by the numerous local optima.

Different from multiple target tracking where the motion of each target is usually independent of the others, the physical links among different limbs impose motion constrains upon them. In other words, the motion of each limb must be spatially coherent with the others, which is reinforced by the kinematic structures of the articulated limbs. We can have an intuitive comparison of these two cases by the configuration space which is the joint motion space of the set of limbs. If the motions of limbs are independent, the configuration space will enjoy a nice property

* Corresponding author. Fax: +1 425 936 7329.
E-mail addresses: ganghua@microsoft.com (G. Hua), yingwu@ece.northwestern.edu (Y. Wu).
[1] Fax: +1 847 491 4455.

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.11.020

that the motion of each limb stays in a manifold which is orthogonal to the manifold corresponding to the other limbs. Thus, independent trackers can be used to track independent multiple targets and the complexity is almost linear w.r.t. the number of targets. However, when the limbs are physically linked, the configuration space will not have such a nice orthogonality and factorization property. Thus, the high dimensionality seems unavoidable. Various approaches have been investigated to alleviate the computation complexity caused by high dimensionality, such as dynamic programming [10,11], annealed sampling [12], partitioned sampling [13,14], eigen-space tracking [15], hybrid Monte Carlo filtering [16], covariance scaled sampling [17], etc., to name a few.

Different from these approaches, in this paper, we propose a novel solution based on a dynamic Markov network [18–20] and a mean field variational analysis. The proposed dynamic Markov network encodes the spatial coherence of different limbs in an undirected graphical model associated with the image observation processes, thus the model serves as a generative model for the articulated motion. We perform Bayesian inference based on a variational mean field approximation, by which tight approximation may be achieved while the computational complexity is significantly reduced. At each time instance, the mean field solution is achieved through Monte Carlo simulation.

To alleviate the problem caused by local optima, we further constrain the variational distribution to be multi-variate Gaussian. Then, we could nicely incorporate a deterministic annealing (DA) scheme into the mean field fixed point iterations to obtain the optimal MAP estimate of the motion. The theoretic foundation of such an variational MAP algorithm, is based on a theorem proven in [21]. The variational MAP algorithm [21], with theoretical guarantee, could achieve better MAP estimate of the motion while only increase the computation linearly.

Related works are discussed in Section 2. In Section 3, we present a decentralized probabilistic representation of the articulated body based on a Markov network. In Section 4, we present the mean field variational method to achieve the Bayesian inference. Due to the multi-modality of the motion posteriors, we use Monte Carlo simulation to implement it. This results in the mean field Monte Carlo (MFMC) algorithm [18–20] in Section 5. We then present the variational MAP algorithm [21] in Section 6. Experimental results are presented and discussed in Section 7. Finally we conclude in Section 8.

2. Related work

There is a substantial literature on articulated motion analysis, and many different approaches have been investigated. For all these methods, three important issues should be addressed: the representations for articulated objects, the computational paradigms, and the means of reducing the computation for the optimal solution.

There can be two typical representations for articulated objects. One employs the joint angles [9,22,14,23], which is in nature a centralized model. While the other uses the collection of the motion of all the limbs, e.g., the cardboard person [24], the decentralized probabilistic model based on Markov network [18,21] which is also used in this paper, the loose-limbed model [25,26], and tree structured model [27,11], to list a few.

Of course, the centralized joint angle representation is non-redundant and reflects the degrees of freedom of the articulated motion directly, while the second one is highly redundant. The centralized representation usually results in a very high dimensional parameterization. Since there are complex motion constraints, it may be possible to learn a lower dimensional manifold to characterize the articulated motion [23,28]. However, the intrinsic dimensionality of the learned manifold may still be quite high. In this case, the motion analysis problem can be posed as an unconstrained optimization in a high dimensional space. On the other hand, if the articulated motion is redundantly described by the individual motion of the subparts, each subpart may be solved individually, and then projected to the constrained space which reinforces the spatial coherence among them. Thus, it corresponds to a constrained optimization problem. By taking advantage of the structure of the configuration space resulted from such a redundant representation, efficient solutions can be found as in this paper.

There are mainly two different computational paradigms for articulated motion analysis: the deterministic approach usually formulates the problem as a parameter estimation problem [29,24,22], and the solution is usually provided by some non-linear optimization methods. While the probabilistic approach formulates it as a Bayesian inference problem [12,18,26], and the solution is provided by recovering the motion posterior sequentially at each time instant. Due to the non-Gaussian densities which commonly exist in a probabilistic formulation [30,31], closed-form implementation of the Bayesian inference is usually intractable and thus it is performed by Monte Carlo simulation. However, both approaches are confronted by the high dimensionality. More specifically, for the deterministic approach, the optimization needs to be performed in a very high dimensional parametric space which is confronted by the numerous local optima. As for the probabilistic approach, the computational cost of a Monte Carlo algorithm may increase exponentially with the dimensionality [32]. Moreover, obtaining the MAP estimate of the motion is also confronted by the same local optima problem as that in the deterministic approach.

Numerous techniques have been proposed to improve the efficiency for the probabilistic approach. For example, a multiple hypothesis tracking algorithm was proposed, which only keeps the salient modes of the motion posteriors for more efficient Monte Carlo simulation [9]. Partitioned sampling is in the spirit of coordinate descent and preforms the sampling in a hierarchical fashion [13,14].

Low dimensional manifold could be learned from the natural hand motion to reduce the dimensionality [23]. In [25,26], the non-parametric belief propagation algorithm [33,34] were applied on the loose-limbed model to achieve the Bayesian inference of the articulated body motion.

It is generally difficult to achieve the optimal MAP estimate since it involves global optimization. It was proven that stochastic simulated annealing (SA) [35–37] methods converge in probability to the global optimum [36]. However, SA algorithms are inherently slow due to the randomized local search strategy. On the other hand, deterministic annealing (DA) [38] utilizes the idea of annealing but it is based on deterministic optimization schemes. This greatly relieves the inefficiency of SA while retaining the benefit of the annealing process. Although global optimality may not be guaranteed by DA, many empirical studies have shown that DA is very likely to achieve global or near global optimal estimate [38].

Different with the previous methods, this paper presents a mean field Monte Carlo (MFMC) algorithm in which a set of low dimensional particle filters interact with one another to collaboratively solve a high dimensional Bayesian inference problem. Moreover, based on a theorem proven in [21], by constraining the variational distribution to be a Gaussian, we could further incorporate a DA scheme into a Gaussian mean field fixed point iterations to pursue the optimal solution.

3. Decentralized probabilistic representation

We denote the motion of each limb by $\mathbf{x}_i$, e.g., it can be the parameters of an affine motion. The motion of an articulated body is the concatenation $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$. Certainly, it is highly redundant. The image observation associated with $\mathbf{x}_k$ is denoted by $\mathbf{z}_k$, which could be the detected edges of the shape contour of the limbs. The collective image observations of the entire articulated body is $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$. An important task is to infer the posterior $p(\mathbf{X}|\mathbf{Z})$.

As shown in Fig. 1, a mixture of undirected and directed graphical model can be used to characterize the generative process. The latent layer is an undirected graph $G_x = \{V, E\}$, representing the spatial coherence among different articulated parts. Obviously, different parts are not independent, and each individual part must be spatially coherent with its neighborhood parts. We denote the neighborhood parts of i by $\mathcal{N}(i)$. Clearly, it is a Markov network. In addition, each individual part is associated with its observation and the conditional likelihood $p(\mathbf{z}_i|\mathbf{x}_i)$ is represented by a directed link.

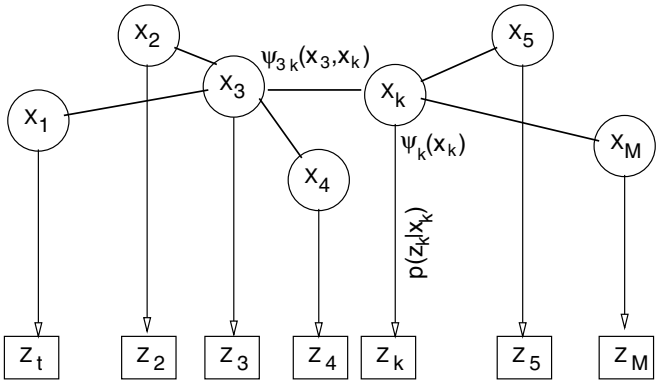

Fig. 1. The Markov network for an articulated body.

Given the undirected graph of $\mathbf{X}$, $p(\mathbf{X})$ can be modelled as a Gibbs distribution and can be factorized as:

$$p(\mathbf{X}) = \frac{1}{Z_c}\prod_{c\in\mathcal{C}}\psi_c(X_c) \tag{1}$$

where c is a clique in the set of cliques $\mathcal{C}$ of the undirected graph, X_c is the set of hidden nodes associated with the clique and $\psi_c(X_c)$ is the probability of this clique, and Z_c is a normalization term or the partition function. Although Z_c is difficult to compute, we do not compute it directly. Instead a Monte Carlo method will be used as shown in later sections. The model accommodates two types of cliques: the first order clique, i.e., $i \in \mathcal{C}^1 = V$, and the second order clique, i.e., $(i,j) \in \mathcal{C}^2 = E$, where $\mathcal{C} = \mathcal{C}^1 \bigcup \mathcal{C}^2$. The associated ψ is denoted by ψ_i and ψ_{ij}, respectively. Thus, Eq. (1) can also be written as:

$$p(\mathbf{X}) = \frac{1}{Z_c}\prod_{(i,j)\in\mathcal{C}^2}\psi_{ij}(\mathbf{x_i},\mathbf{x_j})\prod_{i\in\mathcal{C}^1}\psi_i(\mathbf{x_i}) \tag{2}$$

where $\psi_i(\mathbf{x_i})$ provides a local prior for $\mathbf{x}_i$, and $\psi_{ij}(\mathbf{x_i},\mathbf{x_j})$ presents the spatial coherent constraints between the neighborhood nodes $\mathbf{x}_i$ and $\mathbf{x}_j$. As a specific example, the second order potential $\psi_{ij}(\mathbf{x_i},\mathbf{x_j})$ can be defined as:

$$\psi_{ij}(\mathbf{x_i},\mathbf{x_j}) \propto \mathrm{e}^{-\frac{1}{2}D(\mathbf{x}_i,\mathbf{x}_j)^T\Sigma^{-1}D(\mathbf{x}_i,\mathbf{x}_j)} \tag{3}$$

where $D(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{u}_i(\mathbf{x}_i) - \mathbf{u}_j(\mathbf{x}_j)$, and $\mathbf{u}_i(\mathbf{x}_i)$ and $\mathbf{u}_j(\mathbf{x}_j)$ are shown in Fig. 2. Here, we must emphasize that this zero mean Gaussian prior is a very weak prior, which only captures the connectivity of the neighborhood limbs. The reason we adopt it is that our goal is to analyze arbitrary articulated body motion instead of specific ones. More complex spatial coherence potential functions may be learned for more specific stylized articulated motions.

Given a $\mathbf{x}_i$, its local observation $\mathbf{z}_i$ is independent of the other articulated parts, i.e.,

$$p(\mathbf{Z}|\mathbf{X}) = \prod_{i=1}^{n} p_i(\mathbf{z_i}|\mathbf{x_i}). \tag{4}$$

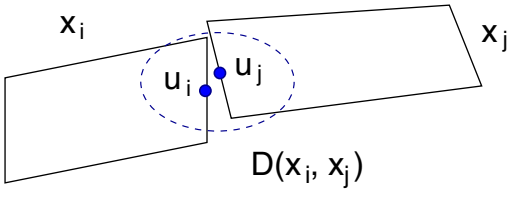

Fig. 2. The spatial coherence constraints of two articulated parts.

Then the problem becomes to infer the posterior $p(\mathbf{x}_i|\mathbf{Z})$. An intuition is that the posterior of $\mathbf{x}_i$ should be affected by three factors: its local prior ψ_i, its local evidence $\mathbf{z}_i$, and the constraints reinforced by its neighborhood through ψ_{ij}. This intuition will become clearer in Section 4. Since the exact analysis of such a model is complicated and involves heavy computation, it is more plausible to have an approximate but efficient solution.

4. Mean field approximation

Variational analysis provides a principled method for approximate Bayesian inference [39–42]. The core idea of variational approximation is to find a variational distribution $Q(\mathbf{X})$ to approximate the posterior distribution $p(\mathbf{X}|\mathbf{Z})$, such that the Kullback–Leibler (KL) divergence of these two distributions is minimized, i.e.,

$$Q^*(\mathbf{X}) = \arg\min_{Q} \mathrm{KL}(Q(\mathbf{X})\|p(\mathbf{X}|\mathbf{Z})) \tag{5}$$

$$= \arg\min_{Q} \int_x Q(\mathbf{X}) \log \frac{Q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Z})} \tag{6}$$

Selecting a good class of variational distributions Q would largely ease the difficulties, but it requires substantial creativity [39]. Here, we adopt a fully factorized form for simplicity:

$$Q(\mathbf{X}) = \prod_i^M Q_i(\mathbf{x}_i) \tag{7}$$

where $Q_i(\mathbf{x}_i)$ only relies on $\mathbf{x}_i$. Then, $H(Q) = \sum_i H(Q_i)$, where $H(Q)$ is the entropy of $Q(\mathbf{X})$, and $H(Q_i)$ is the entropy of $Q(\mathbf{x}_i)$. For each Q_i, the KL-divergence can be written as:

$$\mathrm{KL}(Q_i) = -\sum_{k\neq i} H(Q_k) + \log p(\mathbf{Z}) - H(Q_i) - \int_{x_i} Q_i(\mathbf{x}_i) E_Q[\log p(\mathbf{X},\mathbf{Z})|\mathbf{x}_i] \tag{8}$$

where $E_Q[\cdot|\mathbf{x}_i]$ is the conditional expectation given $\mathbf{x}_i$ w.r.t. $Q(\mathbf{X})$, and $\log p(\mathbf{Z})$ is the data likelihood, which is a constant. To search for a set of Q_i to minimize Eq. (8), since each Q_i is constrained to be a valid p.d.f., we should construct a Lagrangian for each Q_i:

$$L(Q_i) = \mathrm{KL}(Q_i) + \lambda\left(\int_{x_i} Q_i - 1\right) \tag{9}$$

Setting the variation of $L(Q_i)$ w.r.t. Q_i and the derivative of $L(Q_i)$ w.r.t. λ to zeros, we have

$$\begin{cases} -\log Q_i(\mathbf{x}_i) - 1 + E_Q[\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i] + \lambda_i = 0 \\ \int_{\mathbf{x}_i} Q_i(\mathbf{x}_i)\mathrm{d}\mathbf{x}_i - 1 = 0 \end{cases} \tag{10}$$

It is easy to solve the equation set and the solution is a set of fixed point equations, i.e., for each $1 \leqslant i \leqslant M$,

$$Q_i(\mathbf{x}_i) = \frac{1}{Z_i} \mathrm{e}^{E_Q[\log p(\mathbf{X},\mathbf{Z})|\mathbf{x}_i]} \tag{11}$$

where Z_i is the partition function for normalization. The iterative updating of $Q_i(\mathbf{x}_i)$ will monotonically decrease the KL divergence, and eventually reach an equilibrium. These fixed-point equations are called *mean field equations*.

Moreover, the factorization of $p(\mathbf{X})$ in Eq. (2) and $p(\mathbf{Z}|\mathbf{X})$ in Eq. (4) enables further simplification of the mean field equations in Eq. (11). It is easy to show that:

$$Q_i(\mathbf{x}_i) = \frac{1}{Z'_i} p_i(\mathbf{z}_i|\mathbf{x}_i)\psi_i(\mathbf{x}_i)M_i(\mathbf{x}_i) \tag{12}$$

where

$$M_i(\mathbf{x}_i) = \exp\left\{\sum_{k\in\mathcal{N}(i)} \int_{x_k} Q_k(\mathbf{x}_k)\log\psi_{ik}(\mathbf{x}_i,\mathbf{x}_k)\right\} \tag{13}$$

where Z'_i is a constant, and $\mathcal{N}(i)$ is the neighborhood of the subpart i. From Eq. (12), the intuition stated at the end of Section 3 is more pronounced, i.e., the variational belief of a limb $\mathbf{x}_i$ is determined by three factors: the local conditional likelihood $p_i(\mathbf{z}_i|\mathbf{x}_i)$, the local prior $\psi_i(\mathbf{x}_i)$, and the beliefs from the neighborhood limb $\mathbf{x}_{\mathcal{N}(i)}$ (we call it neighborhood prior). This is illustrated in Fig. 3.

Thus,we can treat the term $p_i(\mathbf{z}_i|\mathbf{x}_i)\psi_i(\mathbf{x}_i)$ as an analogue to the local belief, and treat the term $M_i(\mathbf{x}_i)$ as an analogue to the "message" propagated through the nearby subpart of $\mathbf{x}_i$ in the belief propagation algorithm [43], but the computation of $M_i(\mathbf{x}_i)$ here is easier. In addition, we can clearly see from these equations that the computation is significantly reduced by avoiding multi-dimensional integrals, since Eq. (12) involves only single integral.

5. Monte Carlo implementation

5.1. Mean field Monte Carlo (MFMC)

When all the distributions in the Markov network are Gaussian, then we may obtain a closed-form implementation of the fixed point equations. But in visual tracking, the likelihood function $p(\mathbf{z}_i|\mathbf{x}_i)$ is usually non-Gaussian [31] due to the background clutter. This results in that the analytical solution is usually intractable. In this section, we propose a Monte Carlo method to implement the mean

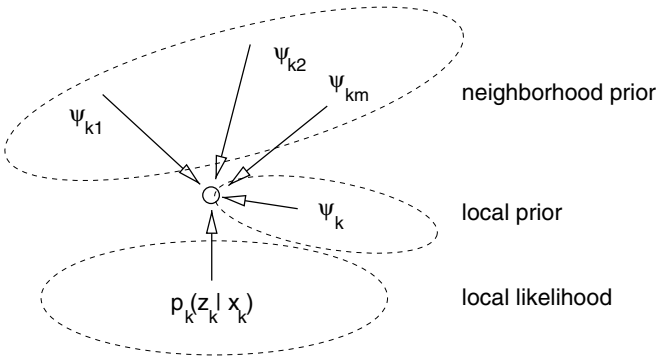

Fig. 3. Three factors affect the updating of $Q(\mathbf{x}_k)$.

field iteration as discussed in Section 4. We call it *mean field Monte Carlo* (MFMC).

Once the mean field iterations converge, then the set of optimal variational distributions $Q_i(\mathbf{x}_i)$, where $i = 1, \ldots, M$, is obtained and can be treated as the optimal approximation to the posterior density $p(\mathbf{x}_i|\mathbf{Z})$.

To make the presentation clear, here we present the mean field updating on one node $\mathbf{x}_i$. We use i and $j \in \mathcal{N}(i)$ to index the node we want to update and the linked neighboring nodes, respectively. In addition, we use k to index the mean field iteration. At the $k-1$th iteration, for each subpart, a set of particle is maintained to represent the variational distribution, i.e.,

$$\begin{cases} Q_i^{k-1}(\mathbf{x}_i) \sim \{s_i^{(n)}(k-1), \pi_i^{(n)}(k-1)\}_{n=1}^N \\ Q_j^{k-1}(\mathbf{x}_j) \sim \{s_j^{(n)}(k-1), \pi_j^{(n)}(k-1)\}_{n=1}^N, j \in \mathcal{N}(i) \end{cases} \tag{14}$$

where s and π denote the sample and the weight, respectively. Then at the next iteration, we perform the following steps according to Eq. (12):

(1) Sampling the local prior $\psi_i(\mathbf{x}_i)$ for $\{s_i^{(n)}(k), \frac{1}{N}\}_{n=1}^N$;
(2) calculating the "message" from j:

$$m_{ij}^{(n)} = \sum_{t=1}^{N} \pi_j^{(n)}(k-1) \log \psi_{ij}(s_i^{(n)}(k), s_j^{(t)}(k-1)), j \in \mathcal{N}(i) \tag{15}$$

(3) Performing observation for each particle $s_i^{(n)}(k)$,

$$w_i^{(n)} = p(z_i|s_i^{(n)}(k)) \tag{16}$$

(4) Re-weighting the particles by:

$$\pi_i^{(n)}(k) = w_i^{(n)} \exp\left\{ \sum_{j \in \mathcal{N}(i)} m_{ij}^{(n)} \right\} \tag{17}$$

and normalize to produce $\{s_i^{(n)}(k), \pi_i^{(n)}(k)\}$.
(5) Performing the same steps for all the nodes in the Markov network according to Eq. (12). Then increasing k for next mean field updating.

Since $\mathbf{x}_i$ describes the motion of one limb, its image observation $\mathbf{z}_i$ should be a function of $\mathbf{x}_i$, i.e., $p(\mathbf{z}_i|\mathbf{x}_i)$ is in fact $p(\mathbf{z}_i(\mathbf{x}_i)|\mathbf{x}_i)$. Since $p(\mathbf{z}_i|\mathbf{x}_i)$ will be used to re-weight the belief (or the posterior density) of $\mathbf{x}_i$, the locations of the particles $\{s_i^{(n)}\}$ will affect the faith of approximating the belief by the set of particles, if the ratio of valid particles is not satisfactory (meaning that a small portion of the particles dominates the re-weighting). To enhance the ratio of valid particles, we use importance sampling technique [44] to place the particles to "better" locations.

The only modification on the above mean field Monte Carlo (MFMC) algorithm is on the first step: instead of sampling the local prior $\psi(\mathbf{x}_i)$ directly to produce $\{s_i^{(n)}, \frac{1}{N}\}_{n=1}^N$, we draw samples $\{s_i^{(n)}, \frac{1}{N}\}_{n=1}^N$ from an importance density $g(\mathbf{x}_i)$. After weight compensation, the set of re-weighted particle is still a properly weighted set for the density $\psi(\mathbf{x}_i)$, i.e.,

$$\psi(\mathbf{x}_i) \sim \left\{ s_i^{(n)}, \frac{\psi(s_i^{(n)})}{g(s_i^{(n)})} \right\}_{n=1}^N \tag{18}$$

The selection of importance density can be flexible, as long as it can provides beneficial information. Here, we give an specific example by using a two-link (where i and j are connected limbs). To generate samples for $\psi_i(\mathbf{x}_i)$, we find the means $\bar{s}_i$ and $\bar{s}_j$ from the two particle sets. After identifying the point $\bar{u}_j$ on $\bar{s}_j$ and the median axis $\bar{L}_i$ of $\bar{s}_i$ (see Fig. 4), we sample $u_i^{(n)}$ from $\mathcal{G}(u_i : \bar{u}_j, \Sigma_u)$, and $L_i^{(n)}$ from $\mathcal{G}(L_i : \bar{L}_i, \Sigma_L)$, where $\mathcal{G}$ represents a Gaussian distribution.

Then the sample $s_i^{(n)}$ is produced by $(L_i^{(n)}, u_j^{(n)})$, and the importance density is:

$$g_j(\mathbf{x}_i) = \mathcal{G}(u_i : \bar{u}_j, \Sigma_u)\mathcal{G}(L_i : \bar{L}_i, \Sigma_L) \tag{19}$$

For limbs which are linked to multiple limbs, we can build one such a Gaussian from each of its neighbors. Then a Gaussian mixture with equal weights for each of the Gaussian components can be constructed to form the importance function, i.e.,

$$g(\mathbf{x}_i) = \frac{1}{K} \sum_{j \in \mathcal{N}(i)} g_j(\mathbf{x}_i) \tag{20}$$

where K is the total number of neighbors of $\mathbf{x}_i$. The use of importance sampling techniques greatly enhances the robustness of the mean field Monte Carlo algorithms.

5.2. Dynamic Markov network and sequential mean field Monte Carlo

Sections 4 and 5.1 describe the mean field approximation and mean field Monte Carlo at one time instance. They can be easily extended for tracking. When considering multiple time instances, the model becomes a dynamic Markov network, as shown in Fig. 5. Denote the collection of observations by $\underline{\mathbf{Z}}_t = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_t\}$.

Tracking algorithms aim at inferring $p(\mathbf{X}_t|\underline{\mathbf{Z}}_t)$ by knowing $p(\mathbf{X}_{t-1}|\underline{\mathbf{Z}}_{t-1})$. Since $\mathbf{X}_t$ consists of a number of articulated limbs, the increase of dimensionality will incur exponential increase of computation. The advantage of mean field approximation is that it decouples different parts, and transforms the problem of exponential complexity to a simpler problem with close to linear complexity. The constraint reinforcement needs some computation as a cost, but it is not significant.

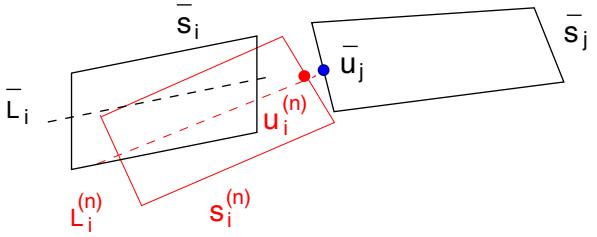

Fig. 4. Importance density.

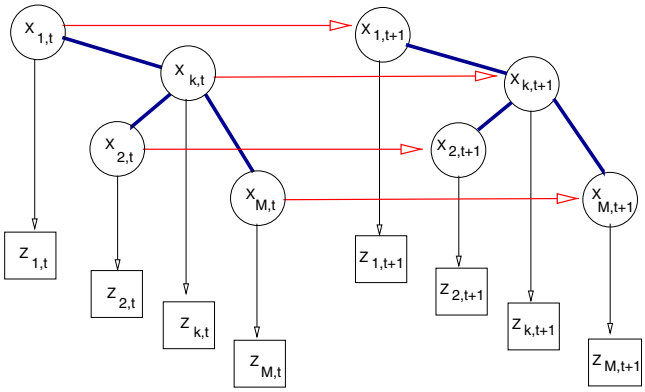

Fig. 5. Dynamic Markov network.

At time instance t, mean field approximation finds a variational distribution $Q_{i,t}(\mathbf{x}_{i,t})$ to approximate $p(\mathbf{x}_{i,t}|\underline{\mathbf{Z}}_t)$ for the ith subpart. The mean field equation can be written as:

$$Q_{i,t}(\mathbf{x}_{i,t}) = \frac{1}{Z_i} p_i(\mathbf{z}_{i,t}|\mathbf{x}_{i,t}) \times \int p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1}) Q_{i,t-1}(\mathbf{x}_{i,t-1}) \mathrm{d}\mathbf{x}_{i,t-1} \times \exp\left\{ \sum_{k \in \mathcal{N}(i)} \int_{x_{k,t}} Q_{k,t}(\mathbf{x}_{k,t}) \log \psi_{ik}(\mathbf{x}_{i,t}, \mathbf{x}_{k,t}) \right\} \quad (21)$$

Comparing Eq. (21) to Eq. (12), we clearly see that the predication density $\int p(\mathbf{x}_{i,t} \mid \mathbf{x}_{i,t-1}) Q_{i,t-1}(\mathbf{x}_{i,t-1}) \mathrm{d}\mathbf{x}_{i,t-1}$ in Eq. (21) of a dynamic Markov network plays the same role as $\psi_i(\mathbf{x}_i)$ in Eq. (12). Thus, at time instance t, the variational belief of the ith subpart is also determined by three factors: the local evidence, the prediction prior from previous time frame, and the belief of the neighborhood subparts.

Therefore, the sequential mean field Monte Carlo can be obtained by modifying the mean field Monte Carlo algorithm in Section 5.1. At the first step, instead of sampling from $\psi_i(\mathbf{x}_i)$, we should sample the prediction prior instead. Suppose at $t-1$, $Q_{i,t-1}(\mathbf{x}_i)$ is represented by:

$$Q_{i,t-1}(\mathbf{x}_i) \sim \{s_{i,t-1}^{(n)}, \pi_{i,t-1}^{(n)}\}_{n=1}^{N}. \quad (22)$$

Then, we can use the following steps to replace the first step in the mean field Monte Carlo algorithm in Section 5.1:

(1a) Re-sampling from $Q_{i,t-1}(\mathbf{x}_i)$ for $\{\widetilde{s}_{i,t-1}^{(n)}, 1\}_{n=1}^{N}$.
(1b) $\forall \widetilde{s}_{i,t-1}^{(n)}$, sampling $s_{i,t}^{(n)}$ from $p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1} = \widetilde{s}_{i,t-1}^{(n)})$.

We have a rough comparison on the computational complexity of the proposed approach with the original CONDENSATION algorithm with joint angle representation. Assume the articulated body consists of M limbs, each of which contributes one DoF, and assume a number of T particles are needed for tracking one limb. In addition, we assume when one more DoF is added, CONDENSATION needs $P \times T$ particles to work. Through our experiments, 10 is reasonable for P. In our mean field Monte Carlo, we denote the number of mean field iteration by K, which is 5 in our experiments. In both methods, the most intensive computation is on calculating image observation, while the extra computation induced by $M_i(\mathbf{x}_i)$ in Eq. (12) is negligible. Thus, the complexity of our method is $O(TKM)$, while CONDENSATION has $O(TP^{M-1})$ which is much larger than the proposed mean field Monte Carlo algorithm when the dimension increases.

In addition, the proposed mean field Monte Carlo (MFMC) algorithm is also different from the partitioned sampling method, although both methods reduce the exponential complexity to close linear complexity. Partitioned sampling takes a hierarchical search strategy which is uni-directional (it may be revised to run back and forth, though), while MFMC is collaborative and iterative, since the fixed point is achieved by the bi-directional interactions among a set of low dimensional particle sets.

6. Variational maximum a posteriori estimation

Since the motion posteriors are usually multi-mode, using the mean estimate may significantly deviate from the MAP estimate and thus could not indicate the true motion. In [21], an annealed variational method called variational MAP, is proposed to approach to the maximum a posteriori estimate. For self-completeness, we briefly summarize the variational MAP algorithm and its theoretical foundations in this section.

We first present a theorem proven in [21], which is about the KL divergence between a Gaussian distribution $q(\mathbf{x})$ and another p.d.f $p(\mathbf{x})$, i.e.,

Theorem 1. *Let $p(\mathbf{x})$, $\mathbf{x}$ is a random vector in $\mathcal{R}^n$, be a bounded, continuous and everywhere positive p.d.f. with the properties*:

- *There exists a unique $\mathbf{x}^* \in \mathcal{R}^n$ such that $p(\mathbf{x}^*) = \sup_{\mathbf{x} \in \mathcal{R}^n} p(\mathbf{x})$*
- *$p(\mathbf{x})$ is proper, i.e., $p(\mathbf{x}) \to \mathbf{0}$ as $\mathbf{x} \to \infty$*
- *The following integrability condition holds*

$$\left| \int_{\mathbf{x}} \exp\left\{ -\frac{\mathbf{x}^T \mathbf{x}}{2} \right\} \log p(\mathbf{x}) \mathrm{d}\mathbf{x} \right| < +\infty \quad (23)$$

Suppose $q(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathcal{I}_n)$ is a Gaussian distribution with zero mean and identity covariance matrix $\mathcal{I}_n$, then denote $q_{\boldsymbol{\sigma}}^{\vec{\mu}}(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}|\vec{\mu}, \boldsymbol{\sigma}^2 \mathcal{I}_n), \mathbf{x} \in \mathcal{R}^n$ as the Gaussian distribution with mean $\vec{\mu}$ and diagonal covariance $\boldsymbol{\sigma}^2 \mathcal{I}_n$. Assume $\vec{\mu}_{\boldsymbol{\sigma}}$ is such that $\mathrm{KL}(q_{\boldsymbol{\sigma}}^{\vec{\mu}_{\boldsymbol{\sigma}}}(\mathbf{x}) \| p(\mathbf{x})) = \inf_{\vec{\mu}} \mathrm{KL}(q_{\boldsymbol{\sigma}}^{\vec{\mu}}(\mathbf{x}) \| p(\mathbf{x}))$, then

$$\lim_{\boldsymbol{\sigma} \to 0} \vec{\mu}_{\boldsymbol{\sigma}} = \mathbf{x}^* \quad (24)$$

Please refer to [21] for the detailed proof of this theorem. This theorem provides the theoretical foundation for pursuing the optimal MAP estimate of the articulated motion. As first revealed in [21], we further constrain the $Q(\mathbf{X})$ to be multi-variate independent Gaussian, i.e.,

$$Q(\mathbf{X}) = \mathcal{G}(\mathbf{X} \ : \ [\vec{\mu}_1^T, \ldots, \vec{\mu}_M^T]^T; \mathrm{diag}[\boldsymbol{\sigma}^2 \mathcal{I}_n, \ldots, \sigma^2 \mathcal{I}_n]) \quad (25)$$

where n is the dimensionality of each $\mathbf{x}_i$ and $\mathcal{I}_n$ is the $n \times n$ identity matrix. First, let $\boldsymbol{\sigma}^2$ be a constant, based on the mean field fixed point equations Eq. (12), which minimizes the KL value in Eq. (8), and following the same strategy of gradient projection [45], we can project the solution to the functional space of the set of Gaussian distributions with fixed covariance $\boldsymbol{\sigma}^2\mathcal{I}_n$ [21], i.e.,

$$\vec{\mu}_i = \frac{1}{\hat{Z}_i}\int_{\mathbf{x}_i} \mathbf{x}_i p_i(\mathbf{z}_i|\mathbf{x}_i)\psi_i(\mathbf{x}_i) \times \exp\left(\sum_{j\in\mathcal{N}(i)}\int_{\mathbf{x}_j} \mathcal{G}(\mathbf{x}_j|\vec{\mu}_j, \sigma^2\mathcal{I}_n)\log\psi_{ij}(\mathbf{x}_i,\mathbf{x}_j)\right) \quad (26)$$

where $\hat{Z}_i$ is again a normalization constant. Then, we can iteratively achieve the optimization over a Gaussian family with fixed covariance by Eq. (26), which we call Gaussian mean field fixed point equations.

According to Theorem 1, we then nicely incorporate a DA scheme to pursue the global maximum of $P(\mathbf{X}|\mathbf{Z})$ [21]. This could be achieved by taking $\boldsymbol{\sigma}^2$ as the temperature T for annealing. We present the variational MAP algorithm [21] as follows:

(1) `Initialization`: $m = 0$; $T_{\max}$ and $T_{\min}$ are very large and very small real positive value, respectively; and $\vec{\boldsymbol{\mu}}_{i,0}$, $i = 1,\ldots,M$ are the initialization mean vectors.
(2) `Annealing`: $m = m + 1$, $T = \frac{T_{\max}}{m}$, then $\boldsymbol{\Sigma}_i = T \times \mathcal{I}_n$; $\vec{\boldsymbol{\mu}}_{i,m} = \vec{\boldsymbol{\mu}}_{i,m-1}$, $i = 1,\ldots,M$; if $T > T_{\min}$, goto Step 3, else goto Step 4.
(3) `Mean field iteration`: Update $\vec{\boldsymbol{\mu}}_{i,m}$ for every $i = 1,\ldots,M$ according to Eq. (26). Iterate this step until convergence. Then jump back to Step 2.
(4) `Result`: $\vec{\boldsymbol{\mu}}^*_i = \vec{\boldsymbol{\mu}}_{i,m}$, $i = 1,\ldots,M$, are the MAP estimation of $P(\mathbf{X}|\mathbf{Z})$.

Note that when T is large, the optimization of the KL is a convex optimization problem [38,21], the iteration of Eq. (26) will surely find the only optimal point. Since the initialization of the fixed point iteration at each T is from the optimization result from the previous annealing step, this DA scheme will guide the whole searching process to the optimal or near optimal point of the real posteriors, as assured by Theorem 1. More detailed discussions can be found in [21].

It is straightforward to extend the variational MAP algorithm to the dynamic Markov network [46], we simply need to replace Eq. (26) with the following equation, i.e.,

$$\vec{\mu}_{i,t} = \frac{1}{\hat{Z}_i} p_i(\mathbf{z}_{i,t}|\mathbf{x}_{i,t}) \times \int \mathbf{x}_{i,t} p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1}) Q_{i,t-1}(\mathbf{x}_{i,t-1})\mathrm{d}\mathbf{x}_{i,t-1} \times \exp\left\{\sum_{k\in\mathcal{N}(i)}\int_{x_{k,t}} \mathcal{G}(\mathbf{x}_{k,t} : \vec{\mu}_{k,t}, \boldsymbol{\sigma}^2\mathcal{I}_n)\log\psi_{ik}(\mathbf{x}_{i,t},\mathbf{x}_{k,t})\right\} \quad (27)$$

which can be derived directly from Eq. (21).

7. Experiments

We performed extensive experiments[2] on articulated body with different DoFs. Impressive results were obtained as reported in this section.

7.1. Experimental setup

Our experiments mainly concern about analyzing the 2D motions. Thus, we adopt a cardboard model where each limb in the articulated body is represented by a planar object, and thus the state of $\mathbf{x}_i$ is the parameters of a 2D affine transform. The motion model $p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})$ is a standard second order constant acceleration model for each limb, which are estimated online based on the estimated motion at each time instant.

The observation model $p(\mathbf{z}_i|\mathbf{x}_i)$ is also an important factor in tracking. We use two types of visual cues: edge and intensity. We adopt the same method in CONDENSATION [31,30] for edge observation, where a set of independent measurement lines were used to measure the likelihood of detected edge points. In addition, since the articulated targets are human body parts and the skin or clothes on the limbs may be similar, we also use the intensity cue and assume that the distribution of the intensity of each limb be a Gaussian distribution. The mean and variance of the Gaussian density is estimated for each individual limb from the manual initialization in the first frame.

7.2. Results of MFMC iteration

To demonstrate that the mean field iterations do converge and are functioning as expected, we collect the intermediate results on the MFMC iterations. The first row and second row of Fig. 6 show that the MFMC iterations at a specific time instant on a 2-part arm and a 3-part finger video sequence, respectively. In both cases, the mean estimates of the motion posteriors at the first five iterations are shown. Before the iteration, the initial status is quite unpleasant. But after a couple of mean field iterations, the estimates settle down on the correct positions as expected. From our experiments, most iterations converge in less than five times. In Section 7.3, we will present the experimental results of exploiting the MFMC algorithm to track various articulated objects. In Section 7.4, we will present the experimental results of articulated body tracking using the variational MAP algorithm.

7.3. Tracking various articulated objects by MFMC

To demonstrate the effectiveness, efficiency and scalability of the MFMC algorithm, we perform experiments on various articulated objects of difference DoFs, including a

[2] The video sequences of all the reported experimental results can be provided upon request.

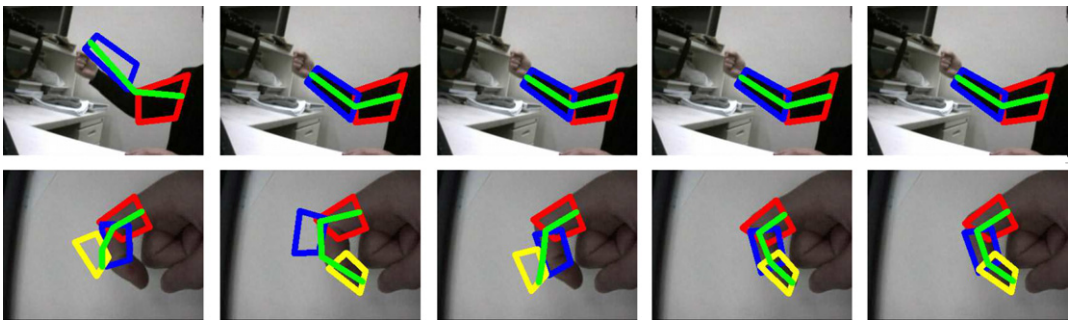

Fig. 6. From left to right, the first five iterations of MFMC at a specific time instant on the (2-part) arm sequence (first row) and (3-part) finger sequence (second row).

2-part arm, 3-part finger, 6-part upper body and 10-part full body. The first test sequence is a 2-part arm, which consists of two limbs: upper arm and lower arm. The sequence consists of 441 frames. The lower arm presents larger motion than upper arm in the testing sequence. The MFMC algorithm perform excellently due to the reinforcement of the spatial coherence constraints. Sample frames are shown in the first row of Fig. 7. We compare the results from MFMC with multiple independent trackers (MiT). Sample result images from MiT are presented in the second row of Fig. 7. Although there are only two limbs, MiT does not produce satisfactory results, since either one has the risks to lose track and there are no other means to get it back except the image observations, and MiT hardly produce plausible results satisfying the spatial coherence constraints.

The second test sequence is on a 3-part finger and consists of 182 frames. As expected, MFMC produce very robust and stable result. Sample frames are shown in Fig. 8.

The third test sequence is on a 6-part upper body, where complex arm motions as well as global movement of the torso and head are presented. The sequence consists of 834 frames. Although the articulation is quite complicated, it does not fail MFMC. Sample frames are shown in Fig. 9. The fourth test sequence is on a 10-part full body, which has 767 frames in total. Arms and legs are the most articulated body parts, and they present significant motion. None of our run of MiT succeeds. Sample results of MiT are shown in the second row of Fig. 10. When MFMC is applied, the tracking result is still very stable unlike MiT. Using the mean estimate, the MFMC algorithm can track the 10-part full body articulation to frame 368. It then loses track because of the heavy multi-modality existed in the motion posteriors where the mean estimate could hardly indicate the true motion. Sample frames are shown in the first row of Fig. 10.

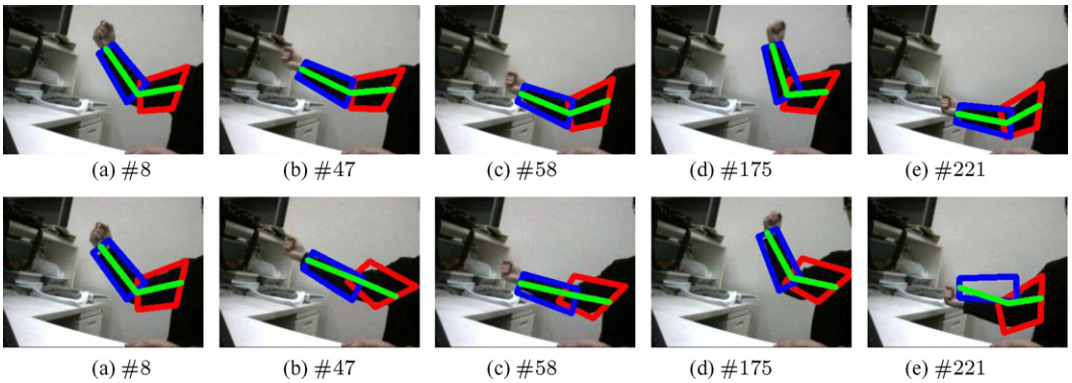

Fig. 7. Tracking 2-part arm: first row, tracking results by MFMC. Second row, tracking results by MiT. Frame numbers are indicated in the bottom of the result image.

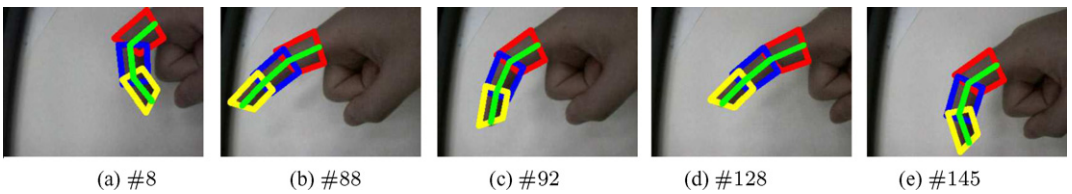

Fig. 8. Mean field Monte Carlo (MFMC): tracking 3-part finger. Frame numbers are indicated in the bottom of the result image.

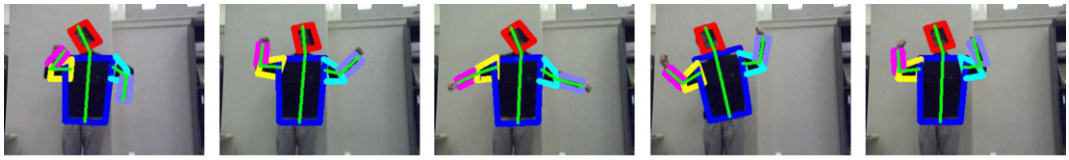

Fig. 9. Mean field Monte Carlo (MFMC): tracking 6-part upper body. Frame numbers are indicated in the bottom of the result image.

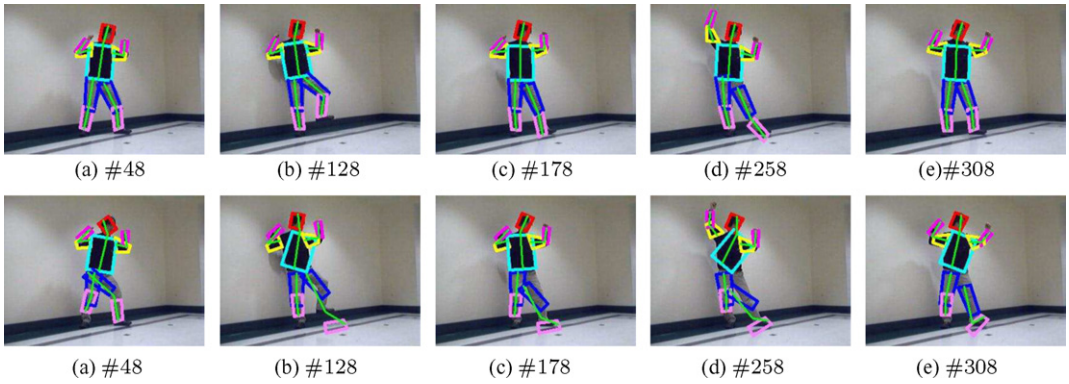

Fig. 10. Tracking 10-part full body: first row, tracking results by MFMC. Second row, tracking results by MiT. Frame numbers are indicated in the bottom of the result image.

7.4. Optimal motion estimate by variational MAP

Although we have obtained satisfactory experimental results using the mean estimate from the MFMC algorithm in Section 7.3, the MAP estimate of the motion posteriors may provide us with better tracking results. In [21], it was shown that on the 10-parts full body sequence presented in Fig. 10, the variational MAP algorithm successfully tracked the articulated body across all the 767 frames. The reason for the better results is that the variational MAP algorithm obtains more accurate MAP estimate of the articulated motion at each time instant. That enables more accurate online estimate of the dynamic model for each limb, which in turn greatly helps to achieve more robust tracking. Please refer to [21] for detailed comparison results. Here, we present some more tracking results on three video sequences, in which a person performs three actions such as "clap", "swing" and "toss". The video sequences have 186, 216, and 335 frames, respectively. These video sequences are more challenging due to the self occlusion between the limbs and torso. The variational MAP algorithm (a Monte Carlo version) [21] obtained

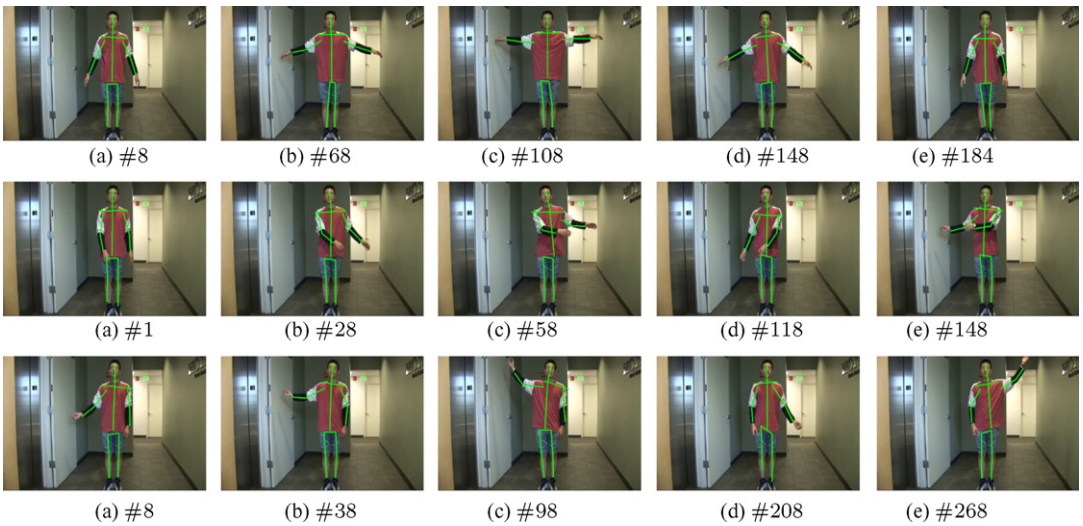

Fig. 11. Tracking 10-part full body by variational MAP: first row, clap sequence; second row, swing sequence; third row, toss sequence. Frame numbers are indicated in the bottom of the result image.

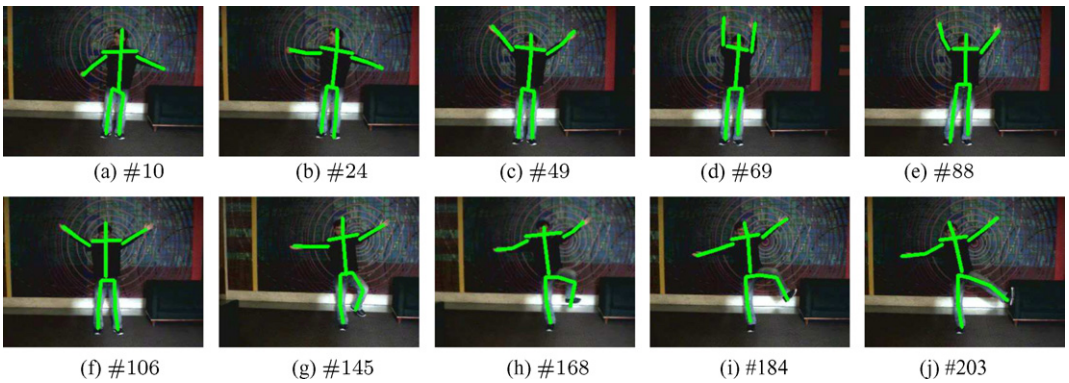

Fig. 12. Tracking 10-part full body. The background present significant clutter. Frame numbers are indicated in the bottom of the result image.

robust results and sample result images are presented in Fig. 11. For clarity, we only overlay the skeleton of the quadrangle shapes of the recovered articulated motion on the sample images.

The last video sequence we have tested is a full-body video sequence of 212 frames, where the background is more cluttered. We presents sample result images in Fig. 12. As we can observe, the background wall behind the moving person is quite cluttered. Although the variational MAP algorithm successfully recovers the motion from the video, the results are less smooth than the results obtained when the background is clean. The reason is that in our current implementation, the motion prior encoded in the hidden layer of the Markov network is a zero mean Gaussian to reinforce the spatial coherence constraints. As we have discussed, this prior is a weak one. Therefore, the image likelihood functions must be reasonable good for the proposed approach to obtain good results. Although we build the image likelihood functions from both edge and intensity cues to make it more robust, the cluttered background may still degrade the quality of the adopted image likelihood function, and thus degrade the quality of the tracking results. Solution to this degradation issue due to clutter may be building better image likelihood functions. For example, when the background is static, we can perform background subtraction first to remove the background clutters. We must emphasize here that none of the experimental results reported here used any background subtraction techniques. When the background is not static, we may build more discriminative image likelihood functions using more image cues such as textures and color distributions. We defer that to our future work.

All the experiments are run on a single processor PC of 2.0 GHz. We do not perform any code optimization. For all these experiments of the MFMC algorithm, the number of mean field iteration is set to 5. For the variational MAP algorithm, we design 6 annealing steps and in the first step of the annealing, we iterate the fixed point equations for 6 times and in the following annealing steps, we iterate the fixed point equations for 3 times. We also design different annealing schemes for different components of the affine state vector since they have different ranges. To be more specific, for the translation components, the annealing starts at $\mathbf{T}_{\max 1} = 8$ and for the scaling components, it starts at $\mathbf{T}_{\max 2} = 0.6$. It is obvious that applying the variational MAP algorithm will increase the computational cost, but the increase is linear compared with the MFMC algorithm. Thus the variational MAP algorithm also achieves close to linear complexity w.r.t the number of limbs.

Table 1
A comparison of the computation of different articulated objects with the MFMC algorithm and the variational MAP algorithm

Algorithm	MFMC				Variational MAP
Experiments	2-part	3-part	6-part	10-part	10-part
Particles/part	200	200	200	200	200
Frame/second	2.02	1.28	0.94	0.56	0.23

7.5. Computation efficiency

To demonstrate the efficiency the MFMC algorithm and the variational MAP algorithm, we present the number of particles for each part and the processing frame rates of both algorithms in Table 1.

As we can observe in the table, with 200 samples for each limb, the processing frame rate decreases almost linearly with the increase of the number of limbs. This indicates that the problem of the exponential increase of computational cost w.r.t the dimensionality has been overcome by the proposed MFMC algorithm. Moreover, we can also observe that the variational MAP algorithm could achieve better results at the expense of more computational cost. But it still achieves linear complexity w.r.t. the number of limbs.

8. Conclusion remarks

Tracking articulated objects is a challenging problem, since the increase of the number of limbs and the physical connection constraints of them would potentially incur high dimensionality, and fail tracking algorithms developed for single target. Thus, algorithms with close to linear

complexity would have much better scalability. In this paper, we propose a decentralized collaborative approach to achieve such a goal. Instead of using the centralized joint angle representation which is irreducible, we adopt a highly redundant representation for articulated body. We represent individual limb by its own motion parameters, but reinforce the spatial coherence constraints among them by a Markov network. Variational analysis is performed for the Bayesian inference on this graphical model. Interestingly, a set of fixed point equations (i.e., the mean field equations) is found, which suggests a collaborative solution to the problem through the iterative interaction among neighboring limbs.

Then a mean field Monte Carlo (MFMC) algorithm is designed to achieve effective computation and a variational MAP algorithm is further adopted to pursue the optimal solution. Extensive experiments demonstrate the effectiveness and scalability of the proposed methods. We also show that the added annealing steps in the variational MAP algorithm does enhance the performance compared with the MFMC algorithm, but that is at the expense of more computation. Nevertheless, we also demonstrate that the computation increase is still linear w.r.t. the number of body parts for the variational MAP algorithm.

Since self-occlusion seems a severe issue for articulated motion, one possible future work is to design collaborative algorithms for solving the occlusion problem. Moreover, since a centralized joint angle representation may facilitate the incorporation of high-order motion constraints, another possible future work would be to design algorithms which combine centralized and decentralized representation together to achieve more efficient and more accurate tracking of the articulated body.

Acknowledgments

This work was performed when Gang Hua was at Northwestern University. It was supported in part by National Science Foundation Grants IIS-0347877, IIS-0308222, Northwestern faculty startup funds for Ying Wu and Walter P. Murphy Fellowship and Richter Fellowship for Gang Hua.

References

[1] Y. Wu, T.S. Huang, Vision-based gesture recognition: A review, in: A. Braffort, R. Gherbi, S. Gibet, J. Richardson, D. Teil (Eds.), Gesture-Based Communication in Human–Computer Interaction, Lecture Notes in Artificial Intelligence, vol. 1739, Springer-Verlag, 1999, pp. 93–104.

[2] L. Bretzner, I. Laptev, T. Lindeberg, Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering, in: Proc. 5th IEEE Int. Conf. Automatic Face and Gesture Recognition, 2002, pp. 423–428.

[3] C. Wren, A. Azarbayejani, T. Darrel, A. Pentland, Pfinder: real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 9 (1997) 780–785.

[4] I. Haritaoglu, D. Harwood, L. Davis, W4: Who? when? where? what? a real time system for detecting and tracking people, in: Proc. IEEE Int. Conf. Face and Gesture Recognition, Nara, Japan, 1998, pp. 222–227.

[5] R. Tanawongsuwan, A. Bobick, Gait recognition from time-normalized joint-angle trajectories in the walking plane, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, 2001, pp. 726–731.

[6] L. Wang, H. Ning, T. Tan, W. Hu, Fusion of static and dynamic body biometrics for gait recognition, in: Proc. IEEE Int. Conf. Computer Vision, 2003, pp. 1449–1454.

[7] Y. Wu, T.S. Huang, View–independent recognition of hand postures, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. II, 2000, pp. 88–94.

[8] Y. Wu, T.S. Huang, Self-Supervised learning for visual tracking and recognition of human hand, in: Proc. AAAI National Conf. Artificial Intelligence, 2000, pp. 243–248.

[9] T.-J. Cham, J.M. Rehg, A multiple hypothesis approach to figure tracking, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Ft. Collins, CO, 1999, pp. 239–245.

[10] Y. Song, X. Feng, P. Perona, Towards detection of human motion, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Hilton Head Island, SC, 2000, pp. 1810–1817.

[11] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, International Journal of Computer Vision 61 (1) (2005) 55–79.

[12] J. Deutscher, A. Blake, I. Reid, Articulated body motion capture by annealed particle filtering, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, 2000, pp. 2126–2133.

[13] J. MacCormick, A. Blake, A probabilistic exclusion principle for tracking multiple objects, in: Proc. IEEE Int. Conf. Computer Vision, Greece, 1999, pp. 572–578.

[14] J. MacCormick, M. Isard, Partitioned sampling, articulated objects, and interface-quality hand tracking, in: Proc. European Conf. Computer Vision, vol. 2, 2000, pp. 3–19.

[15] M. Black, A. Jepson, Eigentracking: Robust matching and tracking of articulated object using a view-based representation, in: Proc. European Conf. Computer Vision, vol. 1, Cambridge, UK, 1996, pp. 343–356.

[16] K. Choo, D. Fleet, People tracking using hybrid Monte Carlo filtering, in: Proc. IEEE Int. Conf. Computer Vision, vol. II, Vancouver, Canada, 2001, pp. 321–328.

[17] C. Sminchisescu, B. Triggs, Estimating articulated human motion with covariance scaled sampling, International Journal of Robotics Research 22 (6) (2003) 371–393.

[18] Y. Wu, G. Hua, T. Yu, Tracking articulated body by dynamic markov network, in: Proc. IEEE Int. Conf. Computer Vision, Nice, Côte d'Azur, France, 2003, pp. 1094–1101.

[19] G. Hua, Y. Wu, T. Yu, Analyzing structured deformable shapes via mean field monte carlo, in: Proc. IEEE Asia Conf. Computer Vision, Jeju Island, Korea, 2004.

[20] G. Hua, Y. Wu, Sequential mean field variational analysis of structured deformable shapes, Computer Vision and Image Understanding 101 (2) (2006) 87–99.

[21] G. Hua, Y. Wu, Variational maximum a posteriori by annealed mean field analysis, IEEE Transaction on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1747–1781.

[22] J.M. Rehg, T. Kanade, Model based tracking of self-occluding articulated objects, in: Proc. Int. Conf. Computer Vision, Cambridge, MA, 1995, pp. 612–617.

[23] Y. Wu, J. Lin, T.S. Huang, Capturing natural hand articulation, in: Proc. IEEE Int. Conf. Computer Vision, vol. II, 2001, pp. 426–432.

[24] S.X. Ju, M.J. Blacky, Y. Yacoobz, Cardboard people: A parameterizedmodel of articulated image motion, in: Proc. Int. Conf. Automatic Face and Gesture Recognition, Killington, Vermont, 1996, pp. 38–44.

[25] L. Sigal, M. Isard, B. Sigelman, M. Black, Attractive people: Assembling loose-limbed models using non-parametric belief propa-

gationAdvances in Neural Information Processing System, 16, MIT Press, 2004.
[26] L. Sigal, S. Bhatia, S. Roth, M. Black, Tracking loose-limbed people, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, 2004, pp. 421–428.
[27] D. Ramanan, D.A. Forsyth, Finding and tracking people from the bottom up, in: Proc. IEEE Conf. Computer Vision and Patter Recognition, vol. 2, Madison, WI, 2003, pp. 467–474.
[28] H. Sidenbladh, M.J. Black, D.J. Fleet, Stochastic tracking of 3d human figures using 2d image motion, in: Proceedings European Conference on Computer VisionLecture Notes in Computer Science, vol. 2, Springer Verlag, Dublin, Ireland, 2000, pp. 702–718.
[29] C. Bregler, J. Malik, Tracking people with twists and exponential map, in: Proc. IEEE Conf, Computer Vision and Pattern Recognition, 1998, pp. 8–15.
[30] A. Blake, M. Isard, Active Contours, Springer-Verlag, 1998.
[31] M. Isard, A. Blake, Contour tracking by stochastic propagation of conditional density, in: Proc. European Conf. Computer Vision, vol. 1, 1996, pp. 343–356.
[32] J. MacCormick, A. Blake, A probabilistic contour discriminant for object localisation, in: Proc. IEEE Int. Conf. Computer Vision, 1998, pp. 390–395.
[33] E.B. Sudderth, A. Ihler, W. Freeman, A. Willsky, Nonparametric belief propagation, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2003, pp. 605–612.
[34] M. Isard, Pampas: Real-valued graphical models for computer vision, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2003, pp. 613–620.
[35] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equations of state calculations by fast computing machine, The Journal of Chemical Physics 21 (1953) 1087–1091.
[36] S. Geman, D. Geman, Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence (1984) 721–741.
[37] S. Kirkpatrick, C.D. Gelatt, M.P.V. Jr., Optimization by simulated annealing, Science 220(4598) (1983) 671–680.
[38] J. Puzicha, T. Hofmann, J.M. Buhmann, Deterministic annealing: Fast physical heuristics for real-time optimization of large systems, in: Proc. 15th IMACS World Conf. Scientific Computation, Modelling and Applied Mathematics, Berlin, 1997.
[39] M. Jordan, Z. Ghahramani, T. Jaakkola, L. Saul, An introduction to variational methods for graphical models, Machine Learning 37 (2000) 183–233.
[40] T.S. Jaakkola, Tutorial on variational approximation methods, MIT AI Lab TR, 2000.
[41] M.J. Beal, Variational algorithms for approximate bayesian inference, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
[42] J.M. Winn, Variational message passing and its application, Ph.D. thesis, Department of Physics, University of Cambridge, 2003.
[43] W. Freeman, E. Pasztor, O. Carmichael, Learning low-level vision, International Journal of Computer Vision 40 (2000) 25–47.
[44] J. Liu, R. Chen, T. Logvinenko, A theoretical framework for sequential importance sampling and resampling, in: A. Doucet, N. de Freitas, N. Gordon (Eds.), Sequential Monte Carlo in Practice, Springer-Verlag, New York, 2000.
[45] J.B. Rosen, The gradient projection method for nonlinear programming. Part I. linear constraints, Journal of the Society for Industrial and Applied Mathematics 8 (1) (1960) 181–217.
[46] G. Hua, Y. Wu, Capturing human body motion from video for perceptual interfaces by sequential variational map, in: Proc. 11th Int. Conf. Human–Computer Interactiona, Las Vegas, 2005, invited.