Capturing Human Body Motion from Video for Perceptual Interfaces by Sequential Variational MAP

Gang Hua, Ying Wu

Department of Electrical& Computer Engineering, Northwestern University
2145 Sheridan Road, Evanston, IL 60208, U.S.A.
{ganghua, yingwu} @ ece.northwestern.edu

Abstract

In this paper, we propose a novel sequential variational maximum a posteriori (MAP) algorithm to recover the articulated human body motion from video for perceptual interfaces. Most probabilistic methods for visual tracking adopt the mean values of the motion posteriors as the estimate. This is due to the general difficulty of the global optimization involved in the MAP estimation. However, the mean estimate is confronted with the tracking failure resulted from the multi-mode motion posteriors. We show, with theoretic guarantee, that the MAP estimate could be asymptotically achieved from a probabilistic variational approach. This new algorithm, namely sequential variational MAP, could recover the human articulation more robustly. It also achieves linear complexity w.r.t. the number of body parts, which greatly relieves the curse-of-dimensionality. Our experimental results demonstrate the effectiveness and efficiency of the proposed algorithm for articulated human body tracking, and its applicability to vision based perceptual interfaces.

1 Introduction

Vision based perceptual interface provides a fully non-invasive way of intelligent human computer interaction (HCI). Such kinds of systems are very important in virtual environment, intelligent home and autonomous video surveillance, etc.. Since gesture and body language play very important roles in our daily communication, they are and should be very important inputs to a vision based perceptual interface.

To utilize human articulation for perceptual interfaces, it is essential to achieve the robust tracking of the articulated motion. There are mainly two approaches: the deterministic approach formulates the problem as a parameter estimation problem (Bregler and Malik 1998, Ju, Blacky and Yacoobz 1996, Rehg and Kanade 1995). The solution is usually provided by some nonlinear optimization techniques; while the probabilistic approach formulates the problem as a Bayesian inference problem (Deutscher, Blake and Reid 2000, Wu, Hua and Yu 2003, Sigal, Bhatia, Roth and Black 2004). And the solution is provided by sequentially recovering the articulated motion posteriors.

The articulated structure of the human body results in a very high dimensional representation. This confronts both approaches, e.g., we need to optimize an objective function of at least 25 degrees of freedom to recover the best estimate of the full human body motion. The computation demand may increase exponentially w.r.t. the dimensionality. Nevertheless, the probabilistic approach became popular due to its flexibility of incorporating useful prior information into the articulated motion tracking system in a principled way.

Because of the convenience in calculation, the mean values of the recovered motion posteriors are often taken as the estimate results (Isard and Blake 1996, Wu et al. 2003, Hua and Wu 2004, Sigal et al. 2004). However, this is inadequate when the posteriors are multi-mode. For example, in contour tracking, the motion posteriors can be non-Gaussian and multi-mode, especially when the background is cluttered (Isard and Blake 1996). Therefore, the mean estimate may significantly deviate from the MAP estimate. And thus it is not able to indicate the true motion.

We propose a novel sequential algorithm to recover the MAP estimate of the motion posteriors. By constraining the mean field variational distribution to be Gaussian, a deterministic annealing scheme can be nicely incorporated into the mean field fix-point iterations. Upon convergence, the mean of the variational Gaussian will be very likely to converge to the MAP estimate. This new algorithm achieves linear complexity w.r.t. the number of body parts,

which greatly relieves the curse-of-dimensionality in the particle filtering based algorithm (Isard and Blake 1996).

Section 2 discusses the related work in the literature; a distributed probabilistic representation of the human articulation is presented in Section 3; then, two theorems of the KL divergence are proved, which are the theoretic foundation of this paper; the details of the sequential variational MAP algorithm is presented in Section 5; various experimental results are demonstrated in Section 6; we conclude the paper with some future work in Section 7.

2 Related work

We briefly discuss the previous work on probabilistic articulated human body tracking in this section.

For probabilistic articulated human body tracking, sequential Monte Carlo algorithm provides a flexible means of Bayesian inference (Isard and Blake 1996), but it also suffers from the exponential increase of the computation demand w.r.t. the dimensionality. This confronts the direct sequential Monte Carlo simulation on a centralized joint angle representation of the human body due to the high dimensionality (Cham and Rehg 1999, Deutscher et al. 2000, MacCormick and Isard 2000, Wu, Lin and Huang 2001). Several techniques were proposed to improve the efficiency, e.g., a multiple hypothesis tracking algorithm was proposed by only keeping the salient modes of the motion posteriors for more efficient Monte Carlo simulation (Cham and Rehg 1999); the partitioned sampling (MacCormick and Isard 2000) algorithm performs the Monte Carlo simulation in a hierarchical way based on the partition of the parameter space; while (Wu et al. 2001) proposed to learn a manifold from the natural hand motion to reduce the dimensionality.

In contrast, a distributed representation models the motion of each body parts individually, but they are subject to the constraints from the neighboring body parts. The representatives are the cardboard people (Ju et al. 1996), the Markov network representation (Wu et al. 2003) and the loose-limbed model (Sigal, Isard, Sigelman and Black 2004), to list a few. In (Wu et al. 2003), an efficient sequential mean field Monte Carlo algorithm (MFMC), which reveals a set of collaborative particle filters, was nicely derived from a mean field variational analysis (Jordan and Weiss 2002). Later, (Sigal, Isard, Sigelman and Black 2004, Sigal, Bhatia, Roth and Black 2004) applied the PAMPAS algorithm (Isard 2003) or the nonparametric belief propagation algorithm (Sudderth, Ihler, Freeman and Willsky 2003), to perform the Bayesian inference on the loose-limbed body model. Both algorithms greatly relieve the curse-of-dimensionality through the efficient Bayesian inference facilitated by the distributed representation.

The algorithms discussed above can recover good approximate inference of the posterior distributions, but they are unable to recover the MAP estimate thus the mean estimates are always taken as the results. This may cause serious tracking failure when the articulated motion posteriors are multi-mode. Based on the theorems proved in Section 4, we propose a sequential variational MAP algorithm, which is able to sequentially recover the MAP estimates of the motion posteriors as well as retain the efficiency in computation.

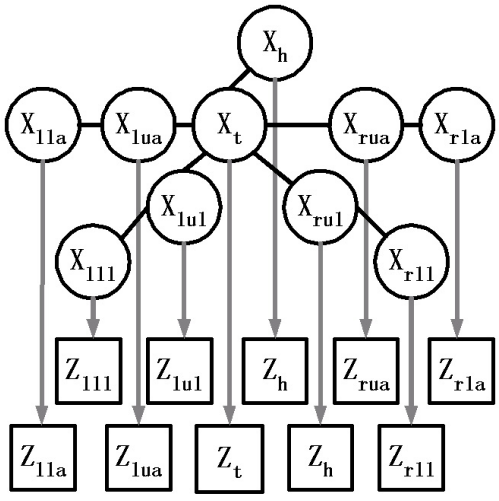

Figure 1: Markov network: a probabilistic distributed representation of human body.

3 Markov network: a probabilistic distributed representation

In this section, we propose a probabilistic distributed representation of the human articulation based on a Markov network similar to that in (Wu et al. 2003). In this representation, the motion of each body parts is modeled individually by a random variable. But each of the random variables is subject to the constraints from the neighboring subparts, e.g., the motion of the lower arm is constrained by the motion of the upper arm, as shown in Figure 1.

Denote L as the set of all the subscripts, then each $X_i, i \in L$ individually models the motion of one of the body part indexed by the subscript, e.g., the subscript "lul" denotes the *left-upper-leg*, "rla" denotes the *right-lower-arm*, etc.. Also, each undirected link in the Markov network represents a potential function $\psi(X_i, X_j)$, which models the motion constraints between two neighboring body parts. And each X_i is associated with an image observation Z_i by a directed link, which represents the image likelihood function $\phi(Z_i \mid X_i)$. Denote $X = \{X_i, i \in L\}$, and $Z = \{Z_i, i \in L\}$, the joint probability of the Markov network is

$$P(X, Z) = \frac{1}{Z_N} \prod_{\{i,j\} \in E} \psi(X_i, X_j) \prod_{i \in L} \phi(Z_i \mid X_i), \tag{1}$$

where Z_N is a normalization constant and E represents the set of all the undirected links. Temporal extension of the Markov network results in the dynamic Markov network to model the human articulation, as shown in Figure 2.

Denote $X_T = \{X_i^T, i \in L\}$ and $Z_T = \{Z_i^T, i \in L\}$ as the set of articulated motion and the set of image observations of all the body parts at time instant T, respectively. Also denote $\underline{Z}_{1:T} = \{Z_1, Z_2, ..., Z_T\}$ as all the image observations up to the current time instant T. Each horizontal directed link in the dynamic Markov network in Figure 2 is associated with the individual motion dynamics of each of the body parts. Thus a fully factorized dynamic model is assumed, i.e.,

$$P(X_{T+1} \mid X_T) = \prod_{i \in L} P(X_i^{T+1} \mid X_i^T). \tag{2}$$

Then, the Bayesian inference here is to sequentially recover the posterior distributions

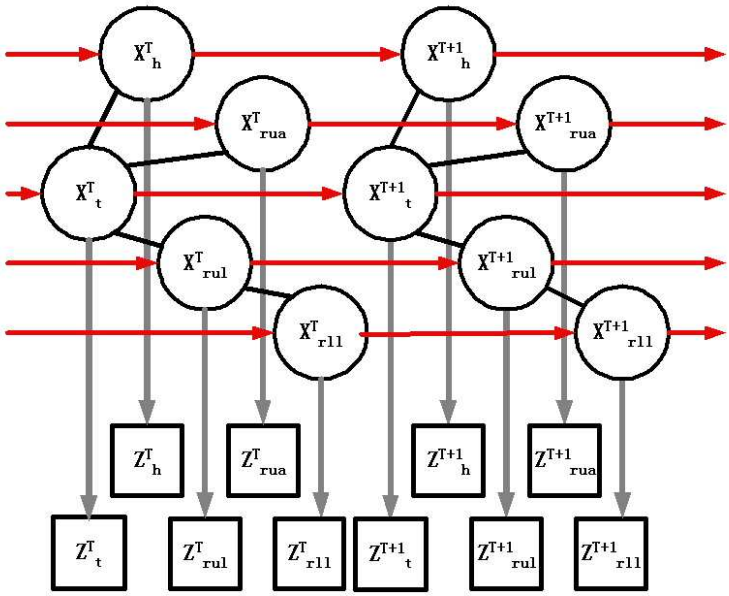

Figure 2: Part of the dynamic Markov network to model the articulated human body motion.

$$\mathrm{P}(\mathrm{X}_{\mathrm{T}} \mid \underline{\mathrm{Z}}_{1:\mathrm{T}}) = \frac{1}{Z_Q} \prod_{i \in L} \phi(Z_i^T \mid X_i^T) \int_{X_{T-1}} \prod_{i \in L} P(X_i^T \mid X_i^{T-1}) \mathrm{P}(\mathrm{X}_{\mathrm{T-1}} \mid \underline{\mathrm{Z}}_{1:\mathrm{T-1}}) dX_{T-1}, \tag{3}$$

where Z_Q is a normalization constant. In Section 5, we will show how to sequentially recover the MAP estimate from an annealed mean field analysis on Equation 3. We will firstly reveal two theorems of the KL divergence between a Gaussian distribution and an arbitrary p.d.f., in Section 4 since they are the theoretic foundation of the proposed algorithm in Section 5.

4 KL divergence between a Gaussian and an arbitrary p.d.f.

The KL divergence between two p.d.f. $q(x)$ and $p(x)$ is defined as

$$KL(q(x) \parallel p(x)) = \int_x q(x) \log \frac{q(x)}{p(x)} dx. \tag{4}$$

It functions as a measurement of the similarity between two distributions. It has the property that it is zero when $q(x)$ and $p(x)$ are equal and is positive otherwise. But it is not a real distance since it is not symmetric, i.e., $KL(q(x) \parallel p(x)) \neq KL(p(x) \parallel q(x))$. We reveal and prove the following two theorems based on the properties of the KL divergence. They provide the theoretic foundation of the sequential variational MAP algorithm in Section 5.

Theorem 1 *For an arbitrary p.d.f.* $p(x), x \in \Re^n$, *which is positive everywhere with an unique global maximum, assuming* $q(x)$ *be a Gaussian distribution with mean* $\vec{\mu}$ *and covariance* Σ, *we have*

$$\lim_{\Sigma \to 0} \arg\min_{\vec{\mu}} KL(q(x) \parallel p(x)) = \arg\max_x p(x) \tag{5}$$

Proof: According to the definition of KL divergence, we have:

$$\begin{aligned}
&\lim_{\Sigma \to 0} \arg\min_{\vec{\mu}} KL(q(x) \parallel p(x)) \\
&= \lim_{\Sigma \to 0} \arg\min_{\vec{\mu}} \left\{ \int_x N(x \mid \vec{\mu}, \Sigma) \log \left(\frac{N(x \mid \vec{\mu}, \Sigma)}{p(x)} \right) dx \right\} \\
&= \lim_{\Sigma \to 0} \arg\min_{\vec{\mu}} \left\{ \int_x N(x \mid \vec{\mu}, \Sigma) \log N(x \mid \vec{\mu}, \Sigma) dx - \int_x N(x \mid \vec{\mu}, \Sigma) \log p(x) dx \right\} \\
&= \lim_{\Sigma \to 0} \arg\min_{\vec{\mu}} \left\{ -\log\left((2\pi e)^n \det(\Sigma) \right) - \int_x N(x \mid \vec{\mu}, \Sigma) \log p(x) dx \right\} \\
&= \lim_{\Sigma \to 0} \arg\min_{\vec{\mu}} \left\{ - \int_x N(x \mid \vec{\mu}, \Sigma) \log p(x) dx \right\} \\
&= \lim_{\Sigma \to 0} \arg\min_{\vec{\mu}} \left\{ - \int_x \delta(x - \vec{\mu}) \log p(x) dx \right\} \\
&= \lim_{\Sigma \to 0} \arg\min_{\vec{\mu}} \{ - \log p(\vec{\mu}) \} \\
&= \arg\max_x p(x) \qquad \blacksquare
\end{aligned}$$

Generally, we also have the following corollary from Theorem 1.

Corollary 1 *The local minima of* $f(\vec{\mu}) = \lim_{\Sigma->0} KL(q(x) \| p(x))$ *have a monotonically one-to-one correspondence to the local maxima of* $p(x)$*, i.e., the global minimum of* $f(\vec{\mu})$ *corresponds to the global maximum of* $p(x)$ *and vice versa.*

Proof: The conclusion is straightforward from the proof of the Theorem 1, since $\log$ function is a monotonically increasing function. ∎

It is worth noting that $f(\vec{\mu})$ may go to infinity, but its topology at infinity can still be characterized by $-\log p(\vec{\mu})$.

5 Sequential variational MAP

The probabilistic inference of Equation 3 by mean field analysis firstly involves the mean field approximation of the motion posteriors at each time instant T, i.e.,

$$P(X_T \mid \underline{Z}_{1:T}) \approx \prod_{i \in L} Q_{i,T}(X_i^T). \tag{6}$$

Embedding Equation 6 at time instant $T-1$ into Equation 3, we have

$$P(X_T \mid \underline{Z}_{1:T}) \approx \frac{1}{Z_Q} \prod_{i \in L} \phi(Z_i^T \mid X_i^T) \int \prod_{i \in L} P(X_i^T \mid X_i^{T-1}) Q_{i,T-1}(X_i^{T-1}) dX_i^{T-1} \tag{7}$$

Then we can construct the following cost function

$$\begin{aligned} J_T(Q) &= \log P(\underline{Z}_{1:T}) - KL\left(\prod_{i \in L} Q_{i,T}(X_i^T) \,\|\, P(X_T \mid \underline{Z}_{1:T}) \right) \\ &= -\sum_{j \in L} H\left(Q_{j,T}\right) + \int Q_{i,T}\left(X_i^T\right) E_Q\left\{\log P\left(X_T, \underline{Z}_{1:T}\right) \mid X_i^T\right\} dX_T \end{aligned}, \tag{8}$$

where $H\left(Q_{j,T}\right)$ is the entropy of the distribution $Q_{j,T}\left(X_i^T\right)$ and

$$E_Q\left\{\log P\left(X_T, \underline{Z}_{1:T}\right) \mid X_i^T\right\} = \oint \prod_{j \in L \setminus i} Q_{j,T}(X_j^T) \log P\left(X_T, \underline{Z}_{1:T}\right) dX_T. \tag{9}$$

We can maximize $J_T(Q)$ to obtain an approximate inference of each $P(X_i^T \mid \underline{Z}_{1:T})$. This is achieved by formulating a Lagrangian multiplier with the constraints that $\int Q_{i,T}\left(X_i^T\right) = 1$. Then using basic calculus of variations, take the variation of the Lagrangian w.r.t. each $Q_{i,T}\left(X_i^T\right)$ and set them to zero, we obtain the following set of mean field fix-point equations

$$Q_{i,T}\left(X_i^T\right) = \frac{1}{Z_S} \exp\left(E_Q\left\{\log P\left(X_T, \underline{Z}_{1:T}\right) \mid X_i^T\right\}\right), \tag{10}$$

where Z_S is the normalization constant. Embedding Equation 7 into Equation 10, we obtain the following sequential mean field fix-point equations.

$$Q_{i,T}\left(X_i^T\right) = \frac{1}{Z_C} \phi\left(Z_i^T \mid X_i^T\right) \int P(X_i^T \mid X_i^{T-1}) Q_{i,T-1}\left(X_i^{T-1}\right) dX_i^{T-1} \exp\left(\sum_{j \in N(i)} Q_{j,T}\left(X_j^T\right) \log \psi\left(X_i^T, X_j^T\right) \right) \tag{11}$$

With the highlight of Theorem 1 and Corollary 1, to pursuit the MAP estimate of the motion posteriors, we further constrain each $Q_{i,T}(X_i^T)$ to be a Gaussian distribution with fixed covariance Σ, i.e.,

$$Q_{i,T}(X_i^T) = \mathrm{N}(X_i^T \mid \vec{\mu}_i^T, \Sigma) \tag{12}$$

Sequential Variational Maximum a Posteriori Algorithm

Input: Unconstrained $Q_{i,T-1}(X_i^{T-1})$ and the MAP estimate $\vec{\mu}_i^{T-1}$ at $T-1$, $i \in L$

Output: Unconstrained $Q_{i,T}(X_i^T)$ and the MAP estimate $\vec{\mu}_i^T$ at T, $i \in L$

1. Initialization: Annealing control parameter $m = 0$; $\mathrm{T}_{\max} = [\mathrm{T}_1^{\max}, \ldots, \mathrm{T}_n^{\max}]$ be very large where the annealing starts and $\mathrm{T}_{\min} = [\mathrm{T}_1^{\min}, \ldots, \mathrm{T}_n^{\min}]$ be very small near zero where the annealing stops; I_n be the $n \times n$ identity matrix; Set $\vec{\mu}_{i,0}^T = \vec{\mu}_i^{T-1}$ as the initialization of the set of mean vectors of the Gaussian distribution.

2. Mean field iteration: Iterate the unconstrained mean field fix-point Equation 10 until convergence to obtain $Q_{i,T}(X_i^T)$, $i \in L$.

3. Annealing: $m = m+1$, $\mathrm{T} = \frac{\mathrm{T}_{\max}}{m}$, then $\sum = \mathrm{TI}_n$; $\vec{\mu}_{i,m}^T = \vec{\mu}_{i,m-1}^T$; if $\mathrm{T} > \mathrm{T}_{\min}$, goto Step 4, else goto Step 5.

4. Gaussian mean field: Update $\vec{\mu}_{i,m}^T$ based on the current value of $\vec{\mu}_{i,m}^T$ and the fixed $\sum$ according to Equation 13. Iterate this step to convergence. Then jump back to Step 3.

5. Result: $\vec{\mu}_i^T = \vec{\mu}_{i,m}^T$, $i \in L$ are the MAP estimation, and $Q_{i,T}(X_i^T)$, $i \in L$, are the optimal unconstrained mean field approximation, of $P(X_i^T \mid \underline{Z}_{1:T})$

Figure 3: the sequential Variational MAP algorithm

Note that maximizing $J_T(Q)$ is equivalent to minimizing $KL\left(\prod_{i \in L} Q_{i,T}\left(X_i^T\right) \| P(X_T, \underline{Z}_{1:T})\right)$. To solve the maximization problem constrained by Equation 12, we follow a similar strategy of gradient projection (Rosen 1960). We firstly relax $Q_{i,T}(X_i^T)$ to be any valid p.d.f., the mean field analysis will result in the fix-point equations in Equation 11. Then, we project the solution to the functional space spanned by the set of Gaussian distributions with fixed covariance $\sum$ by setting the mean $\vec{\mu}_i^T$ to be the expectation of the unconstrained $Q_{i,T}(X_i^T)$, i.e.,

$$\vec{\mu}_i^T = \frac{1}{Z_C} \int X_i^T \phi\left(Z_i^T \mid X_i^T\right) \int P\left(X_i^T \mid X_i^{T-1}\right) Q_{i,T-1}\left(X_i^{T-1}\right) dX_i^{T-1} e^{\sum_{j \in N(i)} \mathrm{N}\left(X_j^T \mid \vec{\mu}_j^T, \Sigma\right) \log \psi\left(X_i^T, X_j^T\right)} dX_i^T . \quad (13)$$

This is the set of fix-point equations to update the Gaussian mean field distribution with fixed covariance $\sum$. Based on this, we can nicely incorporate a deterministic annealing scheme into the Gaussian constrained mean field analysis in Equation 13. This could be achieved by initially setting the elements of the covariance $\sum$ to be very large. Then it will be decreased asymptotically toward zero. At each fixed $\sum$, we iterate Equation 13 until convergence, which uses the converged mean $\vec{\mu}_i^T$ under the previous $\sum$ as the initialization. Then upon convergence of the whole annealed iterations, from Theorem 1 and Corollary 1, the mean of the variational Gaussian distribution will be converged to the global MAP estimate of the posterior $P\left(X_T \mid \underline{Z}_{1:T}\right)$.

Generally, the annealing process of $\sum$ should be carefully designed. For ease of control, we re-enforce $\sum$ to be diagonal, i.e., $\sum = \mathrm{TI}_n$, where $\mathrm{T} = [\mathrm{T}_1, \ldots, \mathrm{T}_n]$ is a n dimensional constant vector and I_n is the $n \times n$ identity

matrix. Then we only need to control n parameters for annealing instead of controlling $\frac{n(n+1)}{2}$ parameters. Note that we must also keep the unconstrained mean field distribution $Q_{i,T-1}(X_i^{T-1})$ at time instant $T-1$ to perform the annealed Gaussian constrained mean field iteration of Equation 13 at the time instant T. We propose the sequential variational MAP algorithm as shown in Figure 3. A hyperbolic decreasing annealing scheme was adopted. It generally achieves good results as shown in our experiments.

6 Experiments

6.1 Recovering human articulation

We implemented the sequential variational MAP algorithm by Monte Carlo simulation to recover the full human body motion from a long video sequence of 767 frames. In the experiment, each body part is represented by a quadrangle shape and tracked in a 6-dimensional probabilistic affine space. The potential function $\psi(X_i, X_j)$ of two connected body parts is modeled by a Gaussian radial basis function. And we use both the visual cues of edge and intensity to construct the image likelihood functions $\phi(Z_i \mid X_i)$. They are all similar to that in (Wu et al. 2003).

The proposed sequential variational MAP algorithm recovers the articulated full-body motion very well across the video sequence. Some of the sample results[1] are in Figure 4. For comparison, we also implemented the mean field Monte Carlo (MFMC) algorithm in (Wu et al. 2003) and multiple independent CONDENSATION trackers (MiCT) to track the human articulation in the same video sequence. Experiments show that the MFMC algorithm failed to track the articulated motion after the 368th frame and the MiCT tracker failed to capture the articulation from the start.

Since different component of the affine motion vector X_i has different range, we designed different annealing scheme for them, e.g., for the translation component, $\mathrm{T}_{\max,i} = 8$, while for the scaling component, $\mathrm{T}_{\max,i} = 0.6$. We design 6 annealing steps and in the first step of the annealing, we iterate the mean field equations for 6 times and in the following annealing steps, we run the mean field fix-point equations for 3 times. This setting is based on the empirical observation that only at the first annealing step that the mean field equations need more iterations to converge. The algorithm can thus run at the speed of 0.2 frames per second. While the MFMC algorithms can run at the speed of 0.6 frames per second where we iterate the mean field fix-point equations for 6 times at each time instant. The proposed sequential variational MAP algorithm does achieve linear complexity w.r.t. the number of body parts, the arguments are similar to that in (Wu et al. 2003) .

6.2 Smart finger mouse

We also applied the proposed algorithm to track the 3 link index finger to demonstrate the potence of developing it to a vision based mouse controller. The articulated motion of the finger is modeled by a Markov network with 3 nodes. We use similar potential functions as well as image observation likelihood functions as in Section 6.1. We define two states of the finger articulation: the key-up state corresponds to when one stretches the index finger to be a near straight line, which we denote as state "0"; and the key-down state corresponds to when the index finger is like a bow shape, which we denote as state "1".

Actually, these two states can be easily characterized by the 2D joint angles of the recovered finger articulation. Denote the joint angle between the distal phalanx link and the middle phalanx as θ_1 and the joint angle between the middle phalanx and the proximal phalanx as θ_2, then the recognition of the two are performed by the following formula, i.e.,

[1] More tracking results of the sequential variational MAP algorithm could be found in the online video at http://www.ece.northwestern.edu/~ganghua/HCIi2005/SVMapArticulate.avi

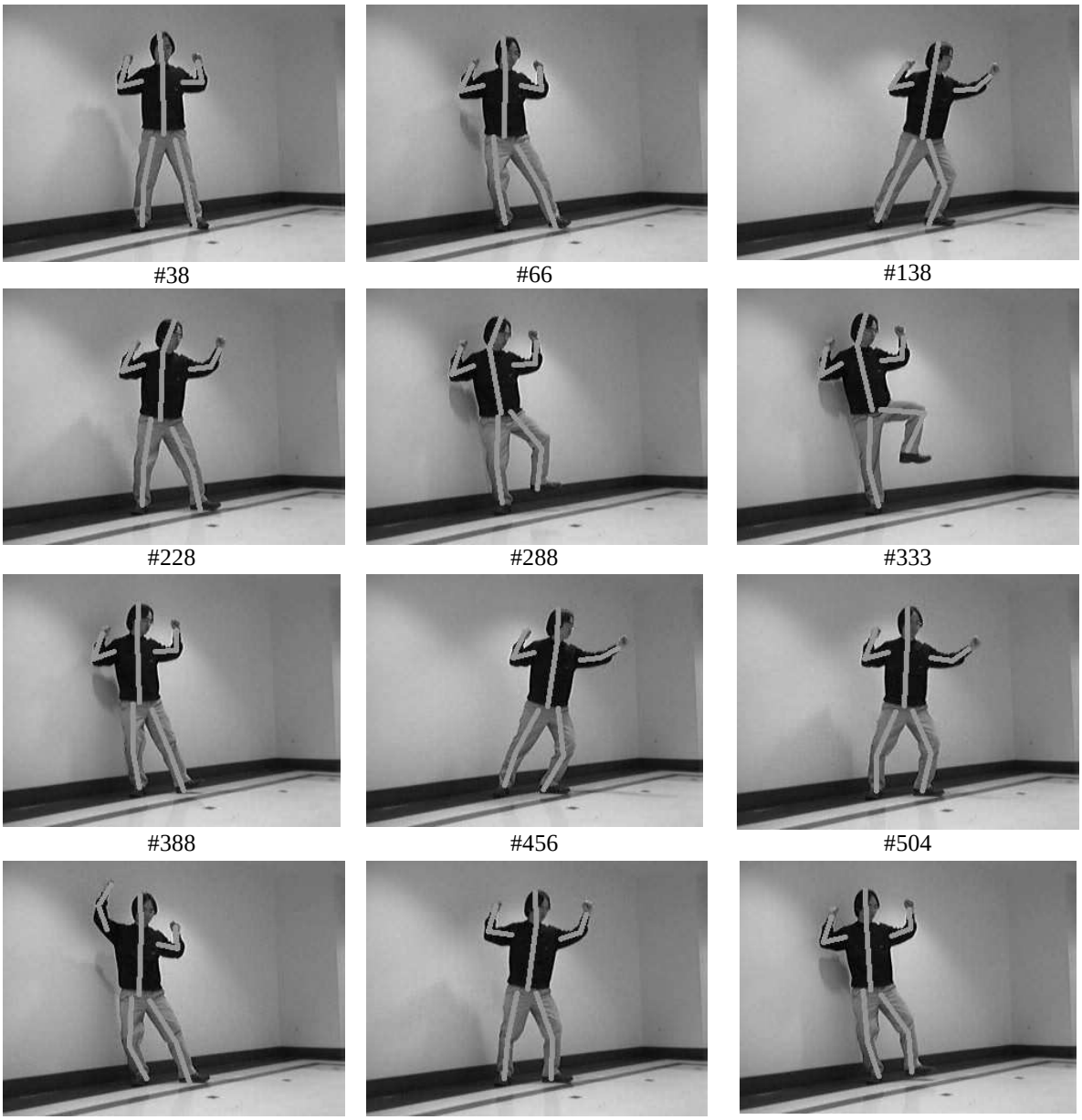

Figure 4: Articulated human body tracking by the sequential variational MAP algorithm

$$S_F = \begin{cases} 0, & \cos\theta_1 \cos\theta_2 > C_T \\ 1, & \cos\theta_1 \cos\theta_2 \le C_T \end{cases}, \tag{14}$$

where C_T is a decision threshold which in our experiment we set it to be 0.9. We present some of the sampling results in Figure 5. We also showed a green cross sign in the image, which corresponds to the boundary point of the joint between the middle phalanx and the proximal phalanx. It functions as the mouse cursor. And we also have shown the recognized finger state in the left top corner of the image as "key down" and "key up". The video sequence has a total 364 frames, our algorithm robustly tracked and recognized the states of the finger articulation across it. Some of the sample results[2] are shown in Figure 5. Without any optimization on the C++ code, the current

[2] More results can be found online at http://www.ece.northwestern.edu/~ganghua/HCII2005/Finger.avi

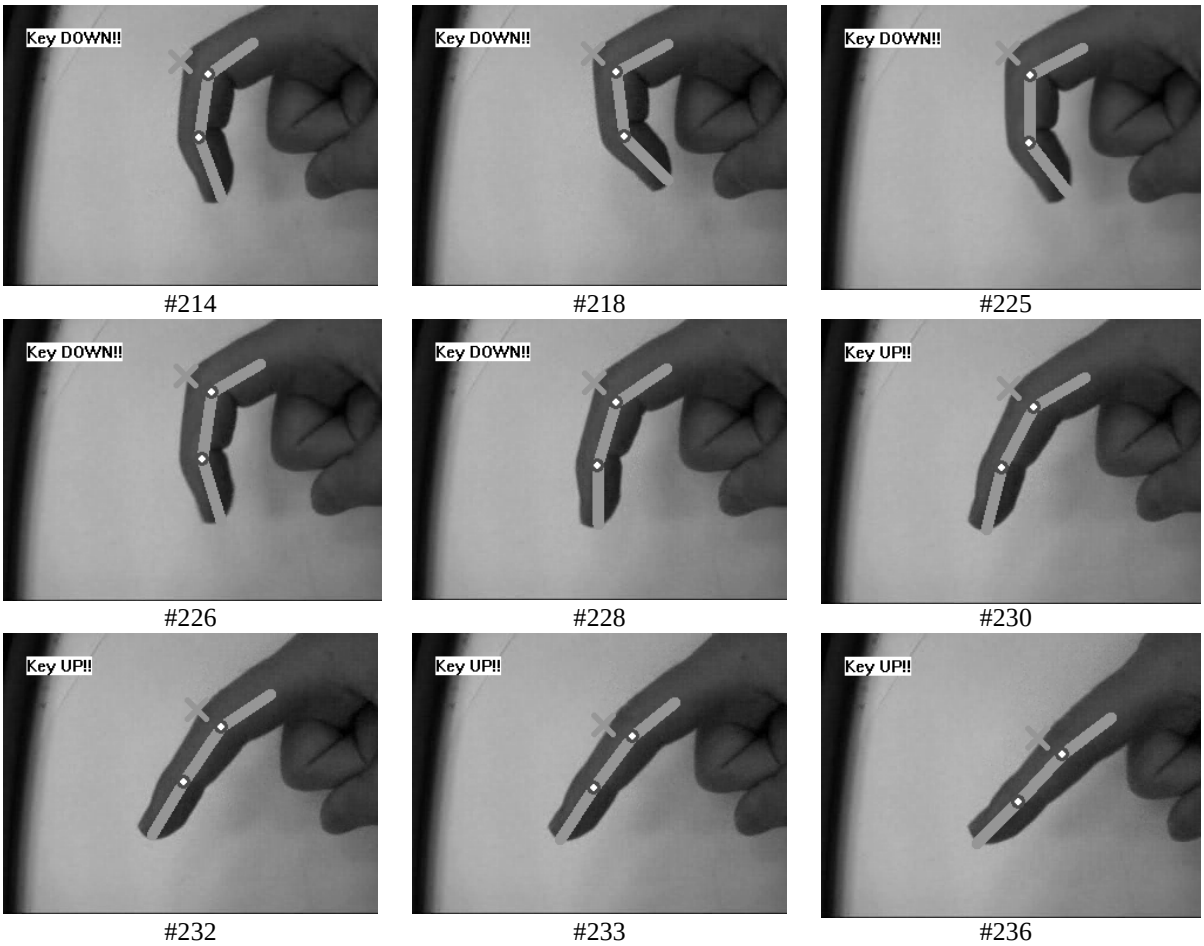

Figure 5: Variational MAP for tracking and recognition of finger motion.

algorithm can run at the speed of 7 frames per second with 50 samples for each body part and 6 annealing steps. The experiments demonstrate the applicability of applying the proposed algorithm for vision based perceptual interface.

In fact, after the finger articulation was robustly recovered and the states were robustly recognized, we can further recognize actions such as “click” and “double-clicks” by using some time series modeling techniques such as hidden Markov model, etc.. Since the motivation of this paper is still focusing on developing algorithms for recovering human articulation more robustly, we defer that part to be our future work.

7 Conclusion and future work

In this paper, we propose a novel sequential variational maximum a posterior algorithm to robustly recover the human articulations from the videos. Different from the previous probabilistic algorithms for tracking articulated motion, which generally take the mean value of the motion posteriors as the estimate, we develop a principled variational approach to sequentially recover the MAP estimate of the articulated motion posteriors. As demonstrated in the experiments, the recovered motion parameters can then be adopted as the input for vision based intelligent human computer interaction.

Our future work include more theoretical investigations on the convergence rate and faster annealing schemes, as that will facilitate to meet the real time requirements for human computer interaction. We will also try to optimize our current implementations of the algorithm and further develop the prototype finger mouse system, e.g., we will seek to develop a principled method for the self-initialization of the proposed sequential variational MAP algorithm.

Acknowledgments

This work was also supported in part by NSF IIS-0347877, IIS-0308222, Northwestern faculty startup funds for Ying Wu and Walter P. Murphy Fellowship for Gang Hua.

Reference

Bregler, C. and Malik, J.: 1998, Tracking people with twists and exponential map, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8–15.

Cham, T.-J. and Rehg, J. M.: 1999, A multiple hypothesis approach to figure tracking, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Ft. Collins, CO, pp. 239–245.

Deutscher, J., Blake, A. and Reid, I.: 2000, Articulated body motion capture by annealed particle filtering, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina.

Hua, G. and Wu, Y.: 2004, Multi-scale visual tracking by sequential belief propagation, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 826–833.

Isard, M.: 2003, PAMPAS: Real-valued graphical models for computer vision, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 613–620.

Isard, M. and Blake, A.: 1996, Contour tracking by stochastic propagation of conditional density, *Proc. European Conference on Computer Vision*, Vol. 1, pp. 343–356.

Jordan, M. and Weiss, Y.: 2002, Graphical models: Probabilistic inference, *The Handbook of Brain Theory and Neural Network*, second edn, MIT Press, pp. 243–266.

Ju, S. X., Blacky, M. J. and Yacoobz, Y.: 1996, Cardboard people: A parameterizedmodel of articulated image motion, *Proc. of International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, pp. 38–44.

MacCormick, J. and Isard, M.: 2000, Partitioned sampling, articulated objects, and interface-quality hand tracking, *Proc. of European Conf. on Computer Vision*, Vol. 2, pp. 3–19.

Rehg, J. M. and Kanade, T.: 1995, Model based tracking of self-occluding articulated objects, *Proc. of International Conference on Computer Vision*, Cambridge, MA, pp. 612–617.

Sigal, L., Bhatia, S., Roth, S. and Black, M.: 2004, Tracking loose-limbed people, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 421–428.

Sigal, L., Isard, M., Sigelman, B. and Black, M.: 2004, Attractive people: Assembling loose-limbed models using non-parametric belief propagation, *Advances in Neural Information Processing System 16*, MIT Press.

Sudderth, E., Ihler, A., Freeman, W. and Willsky, A.: 2003, Nonparametric belief propagation, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 605–612.

Wu, Y., Hua, G. and Yu, T.: 2003, Tracking articulated body by dynamic markov network, *Proc. IEEE International Conference on Computer Vision*, pp. 1094–1101.

Wu, Y., Lin, J. and Huang, T. S.: 2001, Capturing natural hand articulation, *Proc. IEEE Int'l Conference on Computer Vision*, Vol. II, pp. 426–432.