Capturing Hand Articulation in Cluttered Environments by Incorporating Appearance Information in Belief Propagation

Wen-Yan Chang[1,2], Chu-Song Chen[1], and Yi-Ping Hung[1,2]

[1] *Institute of Information Science, Academia Sinica, Taipei, Taiwan*

[2] *Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan*

Email: {*wychang*, *song*}*@iis.sinica.edu.tw*, *hung@csie.ntu.edu.tw*

Abstract

A hybrid framework for tracking in high dimensional space is proposed in this paper. Our framework utilizes the dynamic and static information simultaneously. In our study, two types of static appearance information, explicit and implicit appearances, are introduced to influence the transition model and likelihood in a Bayesian-optimal manner, respectively. A sequential Monte Carlo method is conducted for this hybrid framework and an analysis of statistical distributions is presented for tracking in cluttered environments. To show the performance of the proposed method, we apply it to articulated hand tracking. Our experiments show that the proposed method has better performance than those using either dynamic or static information only.

1. Introduction

Capturing hand articulation is an interesting task in computer vision. However, it is also challenging due to its high degree of freedom (DOF). To avoid searching in high dimensional space, some researches suggested that a set of static images with different states of hand gestures can be collected in advance and a mapping from an image feature space to the hand state can be learned according to this pre-collected static information [2][8][11]. Methods in this category are termed as *appearance-based approaches*. Though high dimensional searching is avoided in these approaches, learning of mapping function and gathering densely distributed static images covering all possible motions are still very demanding for accurate estimation. Hence, studies on tracking in state space directly by using dynamic information from a generative 3D model were proposed [3][7][10][13][14]. Methods in this category are referred to as *dynamic model-driven approaches*.

In a dynamic model-driven approach, state estimation of a dynamic system is formulated for visual tracking. In the past, Stenger *et al.* [10] used the unscented Kalman filter to capture hand motions. To obtain more accurate results, particle filtering was suggested for non-Gaussian state estimation [5] and has been widely used in articulated hand tracking [3][7][14]. Wu *et al.* [14] used particle filtering to recover the motion state in a low-dimensional space by a set of linear manifolds constructed from pre-defined base configurations. Lin *et al.* [7] proposed a stochastic simplex search algorithm by integrating the Nelder-Mead algorithm into particle filtering. Bray *et al.* [3] combined particle filtering with the stochastic meta-descent optimization to find appropriate particles for tracking.

Although state estimation using particle filtering has been shown to be effective for visual tracking, most of the particle filtering-based high DOF hand-tracking methods only used dynamic visual information from previous time steps. Hence, these methods have other limitations. First, as tracking is only initiated from a single state whose appearance is known, the tracking process may easily get trapped in local minimums. Second, existing state estimation methods find it difficult to apply known object appearance information to boost the tracking performance, even when such information is easy to acquire. In addition, most of these works assume that the motions are slow relative to the frame rate of the camera, so that the dynamic transition information from previous time steps plays a decisive role during tracking. Unfortunately, this assumption is not always held and drifting occurs. To overcome these difficulties, a hybrid approach using both static and dynamic information is proposed in this paper. In addition to the dynamic transition information, limited known static appearance information is given as a global guidance for tracking in cluttered environments.

In our approach, the static appearance information is pre-generated or pre-collected. Typically, there are two types of appearance information that can serve as known prior knowledge. The first is *explicit appearances*, which are referred to as some pre-collected observations whose states in state space are known in advance. For a visual tracking task, explicit appearances are some reference images associated with known states in the configuration space. We refer to these known states as “attractors” that guide the tracking in a high dimensional state space. The second type is *implicit appearances*, which are simply some pre-collected observations. For visual tracking, implicit appearances are some reference images pre-gathered, but we do not know exactly which states can generate them. Implicit appearances provide attribute information of objects to assist tracking in a cluttered environment. In this paper, we propose a belief propagation framework that employs both types of appearance information in a Bayesian optimal manner.

2. Review of Particle Filtering

Let the state vector of a target at time t be denote $\boldsymbol{x}_t$, and its observation as z_t. The history of observations from time 1 to t is denoted as $\boldsymbol{Z}_t = \{z_1, \ldots, z_t\}$. The Bayesian formulation of particle filtering is expressed as

$$p(\boldsymbol{x}_t|\boldsymbol{Z}_t) \propto p(z_t|\boldsymbol{x}_t)\cdot p(\boldsymbol{x}_t|\boldsymbol{Z}_{t-1}) = p(z_t|\boldsymbol{x}_t)\int_{\boldsymbol{x}_{t-1}} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\cdot p(\boldsymbol{x}_{t-1}|\boldsymbol{Z}_{t-1}), \quad (1)$$

where a first-order Markov chain is considered, and the likelihood $p(z_t|\boldsymbol{x}_t)$ is the observation model. The Bayesian network (BN) structure of particle filtering is shown in Fig. 1(a). To compute the posterior probability $p(\boldsymbol{x}_t|\boldsymbol{Z}_t)$ for obtaining a Bayesian optimal solution, a closed-form solution with an integral over all possible state values in each iteration is formulated [1][5], but is computationally intractable. In particle filtering, sequential Monte Carlo methods using importance sampling or re-sampling have been adopted to realize the computations [1][4]. The use of importance sampling has been shown to be a powerful strategy for sequential signal processing. However, tracking with particle filtering that depends only on the previous time-step and current frame causes undesirable drifting effects easily, particularly for fast moving target or long image sequences.

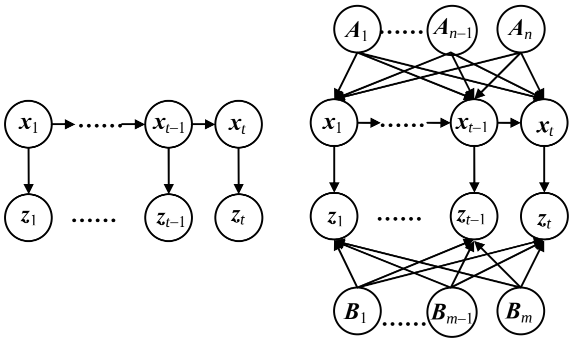
3. Hybrid Approach for Tracking

Unlike classical particle filtering in which tracking employs only sequential dynamic motion transition information, we show how to obtain Bayesian optimal solutions when further prior knowledge of object appearances is imposed.

3.1. Bayesian Formulation

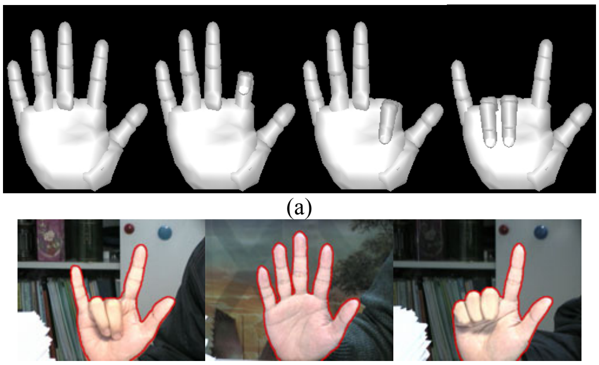
Existing appearance-based approaches require dense samples in the state space. In our approach, however, only a limited number of samples, referred to as "attractors," need to be selected in the state space. Once the attractors are selected, their appearances can be generated by the observation model. Since we know which states can generate these appearances, they are referred to as explicit appearances in this paper.

In our hybrid approach, there are n attractors, $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$, in the state space. These attractors affect the states in $\boldsymbol{X}_t = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$ in such a way that the state $\boldsymbol{x}_t$ is not only influenced by its previous state, $\boldsymbol{x}_{t-1}$, but also by $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$. In addition, suppose m implicit appearances, $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_m$, are also gathered, and the observation $\boldsymbol{Z}_t$ is not only influenced by the state at time t, $\boldsymbol{x}_t$, but also by $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_m$. To illustrate it in a graphical model, the BN of our work in shown in Fig. 1(b). In articulated hand tracking, explicit appearances are generated from the projections of a generic 3D hand model with distinct configuration parameters, and implicit appearances are gathered by capturing hand images with a camera, as shown in Fig. 2.



(a) (b)

Figure 1. The dynamic Bayesian network structure. (a) The BN of classical particle filtering. (b) The BN of our hybrid approach.



(a)

(b)

Figure 2. Some of appearances used in our work. (a) The explicit appearances obtained by rendering a 3D hand model. (b) The implicit appearances in which hand regions were segmented in advance.

With the prior knowledge inherent in the explicit and implicit appearances, the tracking problem we focus on is defined as follows:

> *Given a set of observations from time 1 to t, $\boldsymbol{Z}_t = \{z_1, \ldots, z_t\}$, a set of attractors $\boldsymbol{A} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n\}$, and implicit appearances $\boldsymbol{B} = \{\boldsymbol{B}_1, \ldots, \boldsymbol{B}_m\}$, find the maximal posterior (MAP) estimation of the state $\boldsymbol{x}_t$ according to*

the observations $\boldsymbol{Z}_t$, $\boldsymbol{A}$, and $\boldsymbol{B}$.

The probability of this problem is formulated as

$$p(\boldsymbol{x}_t |\, \boldsymbol{Z}_t,\ \boldsymbol{A},\ \boldsymbol{B}). \tag{2}$$

To resolve (2), let us consider the two basic conditional independencies of the BN shown in Fig. 1(b):

(i) The state at time t, $\boldsymbol{x}_t$, is conditionally independent of the previous states $\boldsymbol{X}_{t-2}$, given the state $\boldsymbol{x}_{t-1}$ and attractors $\boldsymbol{A}$. That is, $p(\boldsymbol{x}_t |\, \boldsymbol{X}_{t-1}, \boldsymbol{A}) = p(\boldsymbol{x}_t |\, \boldsymbol{x}_{t-1}, \boldsymbol{A})$.

(ii) The observation at time t, z_t, is conditionally independent of attractors $\boldsymbol{A}$ and the previous states $\boldsymbol{X}_{t-1}$, given the state $\boldsymbol{x}_t$ and implicit appearances $\boldsymbol{B}$. That is, $p(z_t |\, \boldsymbol{X}_t, \boldsymbol{A}, \boldsymbol{B}) = p(z_t |\, \boldsymbol{x}_t, \boldsymbol{B})$.

In addition, the following three conditional independencies can be derived by the D-separation property [9].

(iii) The observation at time t, z_t, is conditionally independent of the observations $\boldsymbol{Z}_{t-1}$ and the previous states $\boldsymbol{X}_{t-1}$, given the state $\boldsymbol{x}_t$ and implicit appearances $\boldsymbol{B}$. That is, $p(z_t |\, \boldsymbol{X}_t, \boldsymbol{Z}_{t-1}, \boldsymbol{B}) = p(z_t |\, \boldsymbol{x}_t, \boldsymbol{B})$.

(iv) The state at time t, $\boldsymbol{x}_t$, is conditionally independent of the observations $\boldsymbol{Z}_{t-1}$ and implicit appearances $\boldsymbol{B}$, given the previous states $\boldsymbol{X}_{t-1}$. That is, $p(\boldsymbol{x}_t |\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1}, \boldsymbol{B}) = p(\boldsymbol{x}_t |\, \boldsymbol{X}_{t-1})$.

(v) The attractors $\boldsymbol{A}$ are conditionally independent of the observations $\boldsymbol{Z}_t$ and implicit appearances $\boldsymbol{B}$, given the previous states $\boldsymbol{X}_t$. That is, $p(\boldsymbol{A} |\, \boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{B}) = p(\boldsymbol{A} |\, \boldsymbol{X}_t)$.

According to these conditional independencies, the posterior in (2) can be resolved as

$$\begin{aligned}
p(\boldsymbol{x}_t |\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B}) &\propto \int_{\boldsymbol{x}_1..\boldsymbol{x}_{t-1}} p(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B}) \\
&= \int_{\boldsymbol{x}_1..\boldsymbol{x}_{t-1}} p(\boldsymbol{A} |\, \boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{B})\cdot p(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{B}) \\
&= \int_{\boldsymbol{x}_1..\boldsymbol{x}_{t-1}} p(\boldsymbol{A} |\, \boldsymbol{X}_t)\cdot p(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{B}) \qquad (3) \\
&= \int_{\boldsymbol{x}_1..\boldsymbol{x}_{t-1}} p(\boldsymbol{A} |\, \boldsymbol{X}_t)\cdot p(z_t |\, \boldsymbol{x}_t, \boldsymbol{B})\cdot p(\boldsymbol{x}_t |\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1}, \boldsymbol{B})\cdot p(\boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1}, \boldsymbol{B}) \\
&= \int_{\boldsymbol{x}_1..\boldsymbol{x}_{t-1}} p(\boldsymbol{A} |\, \boldsymbol{X}_t)\cdot p(z_t |\, \boldsymbol{x}_t, \boldsymbol{B})\cdot p(\boldsymbol{x}_t |\, \boldsymbol{X}_{t-1})\cdot p(\boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1}, \boldsymbol{B}). \qquad (4)
\end{aligned}$$

In (4), the term $p(\boldsymbol{A} |\, \boldsymbol{X}_t)$ can be rewritten as

$$\begin{aligned}
p(\boldsymbol{A} |\, \boldsymbol{X}_t) &= p(\boldsymbol{A}, \boldsymbol{X}_{t-1}, \boldsymbol{x}_t)/p(\boldsymbol{X}_{t-1}, \boldsymbol{x}_t) \\
&= p(\boldsymbol{x}_t |\, \boldsymbol{X}_{t-1}, \boldsymbol{A})\cdot p(\boldsymbol{A} |\, \boldsymbol{X}_{t-1})/p(\boldsymbol{x}_t |\, \boldsymbol{X}_{t-1}) \\
&= p(\boldsymbol{x}_t |\, \boldsymbol{x}_{t-1}, \boldsymbol{A})\cdot p(\boldsymbol{A} |\, \boldsymbol{X}_{t-1})/p(\boldsymbol{x}_t |\, \boldsymbol{X}_{t-1}). \qquad (5)
\end{aligned}$$

Note that from (3),

$$p(\boldsymbol{x}_{t-1} |\, \boldsymbol{Z}_{t-1}, \boldsymbol{A}, \boldsymbol{B}) \propto \int_{\boldsymbol{x}_1..\boldsymbol{x}_{t-2}} p(\boldsymbol{A} |\, \boldsymbol{X}_{t-1})\cdot p(\boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1}, \boldsymbol{B}).$$

Then, by substituting (5) into (4), we have

$$p(\boldsymbol{x}_t |\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B}) \propto p(z_t |\, \boldsymbol{x}_t, \boldsymbol{B}) \int_{\boldsymbol{x}_{t-1}} p(\boldsymbol{x}_t |\, \boldsymbol{x}_{t-1}, \boldsymbol{A})\cdot p(\boldsymbol{x}_{t-1} |\, \boldsymbol{Z}_{t-1}, \boldsymbol{A}, \boldsymbol{B}). \tag{6}$$

Equation (6) relates the posterior probabilities $p(\boldsymbol{x}_t |\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$ to $p(\boldsymbol{x}_{t-1} |\, \boldsymbol{Z}_{t-1}, \boldsymbol{A}, \boldsymbol{B})$ recursively, which shows how the posterior probabilities propagate given the prior probabilities z_t, $\boldsymbol{A}$, and $\boldsymbol{B}$. Similar to the classical particle filtering, the MAP estimation of $\boldsymbol{x}_t$ can be iteratively obtained from previous time steps. Compared to (1), there are two major distinctions between the classical particle filtering and (6).

- First, in the MAP solution (6), the state transition probability becomes $p(\boldsymbol{x}_t |\, \boldsymbol{x}_{t-1}, \boldsymbol{A})$, instead of $p(\boldsymbol{x}_t |\, \boldsymbol{x}_{t-1})$. This shows how the attractors affect the Bayesian optimal solution in probability propagation.
- Second, the likelihood becomes $p(z_t |\, \boldsymbol{x}_t, \boldsymbol{B})$, instead of $p(z_t |\, \boldsymbol{x}_t)$, which shows how implicit appearances affect the propagation.

Note that unlike some studies based on grid-based filtering [11][12] in which the state space is discrete and consists of a finite number of states, the state space in our method is continuous and unlimited.

3.2. Sequential Monte Carlo Method for Hybrid Approach

In summary of the above, when we treat the MAP estimation of the appearance-incorporated dynamic system as an iterative process, attractors and implicit appearances influence the state transition probability and the likelihood in each iteration, respectively.

The probability propagation solutions of (6) are still computationally infeasible, since the integral over all possible state values is too complex to evaluate. To realize (6), as in classical particle filtering, a set of weighted random samples $\{(\boldsymbol{s}_{t-1}^i, \pi_{t-1}^i),\ i = 1,\dots, N\}$ at time-step $t-1$, where $\boldsymbol{s}_{t-1}^i$ is the sample and π_{t-1}^i is the associated weight, is generated to represent the posterior distribution $p(\boldsymbol{x}_{t-1} |\, \boldsymbol{Z}_{t-1}, \boldsymbol{A}, \boldsymbol{B})$. By using the principle of importance sampling [4], the weights are chosen with $\pi_t \propto p(\boldsymbol{X}_t |\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})/q(\boldsymbol{X}_t |\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$, where $q(\boldsymbol{X}_t |\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$ is an importance distribution (or proposal function) from which it is easier to draw samples than the probability density $p(\boldsymbol{X}_t |\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$.

To derive a useful iteration scheme in which prior appearance information is considered, similar to the derivation of classical particle filtering, we use an importance function that can be factorized as $q(\boldsymbol{X}_t|\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B}) = q(\boldsymbol{x}_t|\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})\cdot q(\boldsymbol{X}_{t-1}|\, \boldsymbol{Z}_{t-1}, \boldsymbol{A}, \boldsymbol{B})$ in which $\boldsymbol{s}_t^i$ can be generated from $q(\boldsymbol{x}_t|\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$ and $\boldsymbol{S}_{t-1}^i = \{ \boldsymbol{s}_1^i,\ldots, \boldsymbol{s}_{t-1}^i;\ i = 1,\ldots,N\}$ can be generated from $q(\boldsymbol{X}_{t-1}|\, \boldsymbol{Z}_{t-1}, \boldsymbol{A}, \boldsymbol{B})$. Then

$$\pi_t \propto p(\boldsymbol{X}_t|\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})/q(\boldsymbol{X}_t|\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$$

$$\propto p(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})/q(\boldsymbol{X}_t|\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B}). \qquad (7)$$

According to (3) and (4), $p(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$

$$= p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A})\cdot p(\boldsymbol{A}|\, \boldsymbol{X}_{t-1})\cdot p(\boldsymbol{z}_t|\, \boldsymbol{x}_t, \boldsymbol{B})\cdot p(\boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1}, \boldsymbol{B})$$

$$\propto p(\boldsymbol{z}_t|\, \boldsymbol{x}_t, \boldsymbol{B})\cdot p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A})\cdot p(\boldsymbol{X}_{t-1}|\, \boldsymbol{Z}_{t-1}, \boldsymbol{A}, \boldsymbol{B}). \qquad (8)$$

By substituting (8) into (7), we have

$$\pi_t \propto \pi_{t-1}\cdot[p(\boldsymbol{z}_t|\, \boldsymbol{x}_t, \boldsymbol{B})\cdot p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A})]/[q(\boldsymbol{x}_t|\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})] \qquad (9)$$

$$= \pi_{t-1}\cdot[p(\boldsymbol{z}_t, \boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B})]/[q(\boldsymbol{x}_t|\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})].$$

The expectation value of π_t, conditional upon $\boldsymbol{X}_{t-1}$, $\boldsymbol{Z}_t$, $\boldsymbol{A}$, and $\boldsymbol{B}$ is

$$\mathrm{E}_{q(\boldsymbol{x}_t|\boldsymbol{X}_{t-1},\boldsymbol{Z}_t,\boldsymbol{A},\boldsymbol{B})}[\pi_t] = \int_{\boldsymbol{x}_t} \pi_t\cdot q(\boldsymbol{x}_t|\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$$

$$\propto \int_{\boldsymbol{x}_t} \pi_{t-1}\cdot p(\boldsymbol{z}_t, \boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B}) = \pi_{t-1}\cdot p(\boldsymbol{z}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B}).$$

Theoretically, an optimal importance distribution can be chosen by minimizing the variance of the importance weights conditional upon $\boldsymbol{X}_{t-1}$, $\boldsymbol{Z}_t$, $\boldsymbol{A}$ and $\boldsymbol{B}$ [4]. Thus,

$$\begin{aligned}&\mathrm{Var}_{q(\boldsymbol{x}_t|\boldsymbol{X}_{t-1},\boldsymbol{Z}_t,\boldsymbol{A},\boldsymbol{B})}[\pi_t]\\&= \mathrm{E}_{q(\boldsymbol{x}_t|\boldsymbol{X}_{t-1},\boldsymbol{Z}_t,\boldsymbol{A},\boldsymbol{B})}[\pi_t^2] - (\mathrm{E}_{q(\boldsymbol{x}_t|\boldsymbol{X}_{t-1},\boldsymbol{Z}_t,\boldsymbol{A},\boldsymbol{B})}[\pi_t])^2\\&\propto (\pi_{t-1})^2[\int_{\boldsymbol{x}_t} \frac{p^2(\boldsymbol{z}_t, \boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B})}{q(\boldsymbol{x}_t \mid \boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})} - p^2(\boldsymbol{z}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B})].\end{aligned}$$

Since $p(\boldsymbol{z}_t, \boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B}) = p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{z}_t, \boldsymbol{A}, \boldsymbol{B})\cdot p(\boldsymbol{z}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B})$, the variance is minimized to zero when the importance distribution is chosen as $q(\boldsymbol{x}_t|\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B}) = p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{z}_t, \boldsymbol{A}, \boldsymbol{B})$. In this case,

$$q(\boldsymbol{x}_t|\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$$

$$= p(\boldsymbol{z}_t|\, \boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B})\cdot p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B})/p(\boldsymbol{z}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B})$$

$$= p(\boldsymbol{z}_t|\, \boldsymbol{x}_t, \boldsymbol{B})\cdot p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A})/p(\boldsymbol{z}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B}). \qquad (10)$$

Substituting (10) into (9), the optimal weight is

$$\pi_t \propto \pi_{t-1}\cdot p(\boldsymbol{z}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}, \boldsymbol{B})$$

$$= \pi_{t-1} \int_{\boldsymbol{x}'_t} p(\boldsymbol{z}_t|\, \boldsymbol{x}'_t, \boldsymbol{B})\cdot p(\boldsymbol{x}'_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}). \qquad (11)$$

However, the optimal importance weight in (11) is difficult to evaluate since an integral is needed. In practice, we choose $q(\boldsymbol{x}_t|\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B}) = p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A})$ instead. This is similar to the common choice for classical particle filtering, $q(\boldsymbol{x}_t|\, \boldsymbol{X}_{t-1}, \boldsymbol{Z}_t) = p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1})$, but the influence of $\boldsymbol{A}$ and $\boldsymbol{B}$ is imposed. In this case, $\pi_t \propto \pi_{t-1}\cdot p(\boldsymbol{z}_t|\, \boldsymbol{x}_t, \boldsymbol{B})$ is obtained.

3.3. Analysis of Distributions

State Transition Distribution

To realize the posterior $p(\boldsymbol{x}_t|\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$ of (6), two distributions, state transition and likelihood, are considered. We set the state transition distribution $p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A})$ as a mixture of $p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_t|\, \boldsymbol{A}_i)$, where $i = 1, \ldots, n$.

$$p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A}) = \alpha_0\cdot p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}) + \Sigma_{i=1,\ldots,n}\alpha_i\cdot p(\boldsymbol{x}_t|\, \boldsymbol{A}_i), \qquad (12)$$

where $\Sigma_{i=0,\ldots,n}\, \alpha_i = 1$.

In (12), $p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1})$ is the probability from the previous state, and $p(\boldsymbol{x}_t|\, \boldsymbol{A})$ is the probability from state similarity. By substituting (12) into (6), we have

$$p(\boldsymbol{x}_t|\, \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B}) \propto p(\boldsymbol{z}_t|\, \boldsymbol{x}_t, \boldsymbol{B})\cdot[\Sigma_{i=1,\ldots,n}\alpha_i\cdot p(\boldsymbol{x}_t|\, \boldsymbol{A}_i) + \int_{\boldsymbol{x}_{t-1}} \alpha_0\cdot p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1})\cdot p(\boldsymbol{x}_{t-1}|\, \boldsymbol{Z}_{t-1}, \boldsymbol{A}, \boldsymbol{B})]. \qquad (13)$$

When α_0 is set to one and the implicit appearances $\boldsymbol{B}$ are disabled, our method degenerates to classical particle filtering in which only dynamic propagation information is used. On the other hand, if α_0 is set to zero, only static information is used and our method degenerates to a pure appearance-based approach. In our implementation, the probabilities $p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_t|\, \boldsymbol{A}_i)$ are modeled by Gaussian distributions,

$$p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}) \sim N(\boldsymbol{x}_{t-1}, \Sigma_1) \text{ and } p(\boldsymbol{x}_t|\, \boldsymbol{A}_i) \sim N(\boldsymbol{A}_i, \Sigma_2),$$

respectively, where Σ_1 and Σ_2 are diagonal covariance matrices.

In our approach, a sample set $\{\boldsymbol{s}_t^i,\ i = 1,\ldots, N\}$ is randomly selected and generated from $\{\boldsymbol{s}_{t-1}^i\}$ via the transition model $p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A})$. Since $p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1}, \boldsymbol{A})$ is defined as a mixture distribution of $p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_t|\, \boldsymbol{A})$, we generate particles for both of them. Let $\{\hat{\boldsymbol{s}}_t^k,\ k = 1,\ldots, M_1\}$ and $\{\boldsymbol{a}^j,\ j = 1,\ldots, M_2\}$ be the sets of samples generated from $p(\boldsymbol{x}_t|\, \boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_t|\, \boldsymbol{A})$, respectively, where $M_1+M_2 = N$. To draw the mixture distribution in (13), M_1 is set as $\alpha_0\cdot N$ and M_2 is set as $(1-\alpha_0)\cdot N$, and the sample set at time-step t $\{\boldsymbol{s}_t^i, i = 1,\ldots, N\}$ is $\{\hat{\boldsymbol{s}}_t^k\}\cup\{\boldsymbol{a}^j\}$.

Likelihood Model

The likelihood considered in our approach is the probability $p(z_t| \boldsymbol{x}_t= \boldsymbol{s}_t^i, \boldsymbol{B})$, which is based on both the state $\boldsymbol{x}_t$ and the implicit appearances $\boldsymbol{B}$. To measure the likelihood, we employ silhouette, edge, and texture information and it is defined as

$$p(z_t| \boldsymbol{x}_t, \boldsymbol{B}) \propto p_{silhouette}(z_t| \boldsymbol{x}_t, \boldsymbol{B})\cdot p_{edge}(z_t| \boldsymbol{x}_t, \boldsymbol{B}). \quad (14)$$

To compute both $p_{silhouette}(z_t| \boldsymbol{x}_t = \boldsymbol{s}_t^i, \boldsymbol{B})$ and $p_{edge}(z_t| \boldsymbol{x}_t = \boldsymbol{s}_t^i, \boldsymbol{B})$, a 3D model corresponding to $\boldsymbol{x}_t$ is projected onto the image plane, so that a $w\times h$ binary silhouette image, $\boldsymbol{I}_t$, is obtained, where $\boldsymbol{I}_t(i, j) = 1$ means the pixel (i, j) belongs to the silhouette projected; otherwise $\boldsymbol{I}_t(i, j) = 0$, for all $i = 1,\ldots,w$ and $j = 1,\ldots,h$. For $p_{silhouette}(z_t| \boldsymbol{x}_t, \boldsymbol{B})$, we compute the area difference E between $\boldsymbol{I}_t$ and z_t as

$$E = \sum_{0<i\le w,0<j\le h}|\boldsymbol{I}_t(i,j) - \boldsymbol{C_B}(z_t(i,j))|, \quad (15)$$

where $\boldsymbol{C_B}$ is a pixel-wise foreground classifier based on implicit appearances $\boldsymbol{B}$. The output value of $\boldsymbol{C_B}(z_t(i,j))$ is set as one if $z_t(i,j)$ is a foreground-color pixel, and zero if $z_t(i,j)$ is a background pixel.

To obtain $\boldsymbol{C_B}$, each implicit appearance in $\boldsymbol{B}$ is segmented into two regions, foreground and background, and $\boldsymbol{B}$ serves as training set to construct a foreground classifier. In our work, the pixel-wise binary classifier $\boldsymbol{C_B}$ is constructed by using the method in [6]. We then set

$$p_{silhouette}(z_t| \boldsymbol{x}_t, \boldsymbol{B}) \propto \exp(-E^2/2\sigma_1^2), \quad (16)$$

where σ_1 is a variance constant.

To estimate $p_{edge}(z_t| \boldsymbol{x}_t, \boldsymbol{B})$, the directed Chamfer distance (DCD) is used, which is relatively robust against small translations, rotations, and deformations of edge images, and has been successfully applied to object recognition and contour alignment. In essence, an edge image can be represented as a set of points corresponding to edge pixel locations. With the implicit appearances, a weighted DCD is formulated in our study. Given two sets of edge points, $\boldsymbol{U}$ and $\boldsymbol{V}$, obtained from the contour of $\boldsymbol{I}_t$ and the Canny edges of z_t respectively, the foreground-color weighted DCD, D_w, is defined as

$$D_w = \frac{1}{\#(\boldsymbol{U})}\sum_{u\in U}\min_{v\in V}\frac{\| u - v \|}{\boldsymbol{C_B}(v) + \delta}, \quad (17)$$

where $\#(\boldsymbol{U})$ is the cardinality of the set $\boldsymbol{U}$, $\|u-v\|$ denotes the Euclidean distance between two pixel locations u and v, and δ is a small positive constant. The likelihood of edge is defined as

$$p_{edge}(z_t| \boldsymbol{x}_t, \boldsymbol{B}) \propto \exp(-D_w^2/2\sigma_2^2), \quad (18)$$

where σ_2 is another variance constant.

The weight π_t^i of each sample in $\{\boldsymbol{s}_t^i\}$ is calculated from the observation distribution $p(z_t| \boldsymbol{x}_t = \boldsymbol{s}_t^i, \boldsymbol{B})$. Thus, the desired posterior distribution $p(\boldsymbol{x}_t| \boldsymbol{Z}_t, \boldsymbol{A}, \boldsymbol{B})$ can be represented by the set of weighted samples $\{(\boldsymbol{s}_t^i, \pi_t^i)\}$. The state $\boldsymbol{x}_t$ at time-step t is estimated by

$$\boldsymbol{x}_t^* = \boldsymbol{s}_t^*, \text{ where } \pi_t^* = \max(\pi_t^i). \quad (19)$$

Instead of tracking objects with fixed models, an adaptive strategy can be adopted by adding the segmented z_t into implicit appearances $\boldsymbol{B}$ in our method and new classifiers $\boldsymbol{C_B}$ are sequentially trained. With the incrementally updated implicit appearances $\boldsymbol{B}$, an effective likelihood estimation can be achieved.

4. Articulated Hand Tracking

In our work, a 3D hand model that has 22 DOFs is used for hand tracking, where each finger has 4 DOFs and the palm has 2 DOFs of rotation. To construct the architecture of the hybrid approach, only a limited number of the attractors are needed, since sequential motion-transition information is also available. Currently, we simply select the attractors by uniformly distributing them in the state space: three attractors for each finger and nine attractors for global hand motion by bending the finger and rotating the palm to different levels. The first three pictures in Fig. 2(a) show an example of the attractors of the index finger. In addition, only fourteen implicit appearances with distinct poses are collected in different cluttered environments.

In our framework, all the attractors affect the state to be estimated with certain probabilities. Nevertheless, for efficiency, we use the most significant K attractors, which have the largest probability values of $p(z_t| \boldsymbol{x}_t=\boldsymbol{A}_i, \boldsymbol{B})$ instead. As the attractors far away from the current state have small probabilities that can be neglected. However, it is also time-consuming to evaluate the values of $p(z_t| \boldsymbol{x}_t=\boldsymbol{A}_i, \boldsymbol{B})$ for all i. To choose the most significant K attractors in an efficient way, a complete graph is constructed in the state space, on which the attractors are the nodes and the edge cost is the distance between nodes in the state space. Instead of computing $p(z_t| \boldsymbol{x}_t=\boldsymbol{A}_i, \boldsymbol{B})$ for all i, only a limited number of attractors, which have the minimum edge costs to the attractors selected in time $t-1$ are evaluated. By so doing, our method becomes more efficient.

5. Experimental Results

In our experiments, we compare the proposed method with both the classical particle filtering approach and the pure appearance-based approach. Image sequences in cluttered environments with a large range of motions are captured. Twenty particles (i.e., $N = 20$) and α_i is set as $1/(K+1)$, for $i = 0,\ldots, K$.

In Fig. 3, it shows the tracking results with 14-DOF by bending three fingers and rotating the palm. In this experiment, K is set as two. There are $9\times3^3 = 243$ attractors, where 3 attractors are for each finger and 9 attractors are for global hand rotation. Part of the input images is shown in Fig. 3(a). The tracking results of classical particle filtering, the pure appearance-based approach, and the proposed hybrid approach are shown in Figs. 3(b), 3(c), and 3(d), respectively. From these results, it can be seen that classical particle filtering drifts when a large range of motions occur as shown in Fig. 3(b). In addition, tracking is not very accurate by using pure appearance information when the appearance corresponding to the current motion state has not been collected in advance, as shown in Fig. 3(c). In contrast to these results, a better articulated motion can be recovered by using the proposed hybrid approach as shown in Fig. 3(d).

More experiments with different DOFs are performed as shown in Figs. 4, 5, and 6. In Fig. 4, it shows the tracking results with 20-DOF by bending five fingers, where 3^5 attractors are used. A four-finger tracking with 3^4 attractors is performed and its results are shown in Fig. 5. In these two experiments, the value of K is set as two. Finally, Fig. 6 shows the results of tracking global hand motion with five-finger articulation, where the DOF is 22 with 2,187 attractors, and K is set as one. From these experiments, one can see that our approach is an effective one for articulated hand tracking. It can also be observed that our approach performs better than classical particle filtering in which only a motion transition model is used, and performs also better when only appearance information is used.

6. Conclusions

In this paper, we present a statistically optimized framework for model-based tracking that incorporates static appearance information into a dynamic belief network. In our approach, two types of appearance information, explicit and implicit appearances, are introduced. We have derived that, in probability propagation of the Bayesian optimal solution, explicit and implicit appearances influence the transition model and likelihood, respectively. A sequential Monte Carlo method has been provided to realize the computation. With these two types of appearances, the proposed method can recover rapid motion in cluttered environments and refine the mis-tracking that is difficult to be handled by classical particle filtering. In addition, unlike pure appearance-based approaches for articulated hand tracking, the proposed method can handle unseen images by taking advantage of the motion transition model, since it is possible to recover motion states even if they are not pre-gathered in the appearance database. In our experiments, promising tracking results are obtained by using the proposed method.

Acknowledgment: This work was supported in part under Grants NSC 93-2213-E-001-011 and 93-EC-17-A-02-S1-032.

References

[1] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. on Signal Processing*, pp. 174-188, Vol. 50, No. 2, 2002.

[2] V. Athitsos and S. Sclaroff, "Estimating 3D Hand Pose from a Cluttered Image," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 432-439, Vol. 2, 2003.

[3] M. Bray, E. Koller-Meier, and L. Van Gool, "Smart particle filtering for 3D hand tracking," *Proc. IEEE Int. Conf. Automatic Face & Gesture Recognition*, pp. 675-680, 2004.

[4] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comput.*, Vol. 10, No. 3, pp. 197-208, 2000.

[5] M. Isard and A. Blake, "CONDENSATION-conditional density propagation for visual tracking," *Int. J. of Computer Vision,* pp. 5-28, Vol. 29, No. 1, 1998.

[6] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Int. J. of Computer Vision*, pp. 81-96, Vol. 46, No. 1, 2002.

[7] J. Y. Lin, Y. Wu, and T. S. Huang, "3D model-based hand tracking using stochastic direct search method," *Proc. IEEE Int. Conf. Automatic Face & Gesture Recognition*, pp. 693-698, 2004.

[8] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff, "3D hand pose reconstruction using specialized mappings," *Proc. IEEE Int. Conf. Computer Vision*, pp. 378-385, Vol. 1, 2001.

[9] S. Russell and P. Norvig, *Artificial intelligence- A modern approach*, Prentice-Hall Inc., 1995.

[10] B. Stenger, P. R. S. Mendonca, and R. Cipolla, "Model-based 3D tracking of an articulated hand," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 310-315, Vol. 2, 2001.

[11] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Filtering using a tree-based estimator," *Proc. IEEE Int. Conf. Computer Vision*, pp. 1063-1070, Vol. 2, 2003.

[12] K. Toyama and A. Blake, "Probabilistic tracking in a metric space," *Proc. IEEE Int. Conf. Computer Vision*, pp. 50-57, Vol. 2, 2001.

[13] Y. Wu and T. S. Huang, "Capturing articulated human motion: a divide-and-conquer approach," *Proc. IEEE Int. Conf. Computer Vision*, pp. 606-611, 1999.

[14] Y. Wu, J. Y. Lin, and T. S. Huang, "Capturing natural hand articulation," *Proc. IEEE Int. Conf. Computer Vision*, pp. 426-432, Vol. 2, 2001.

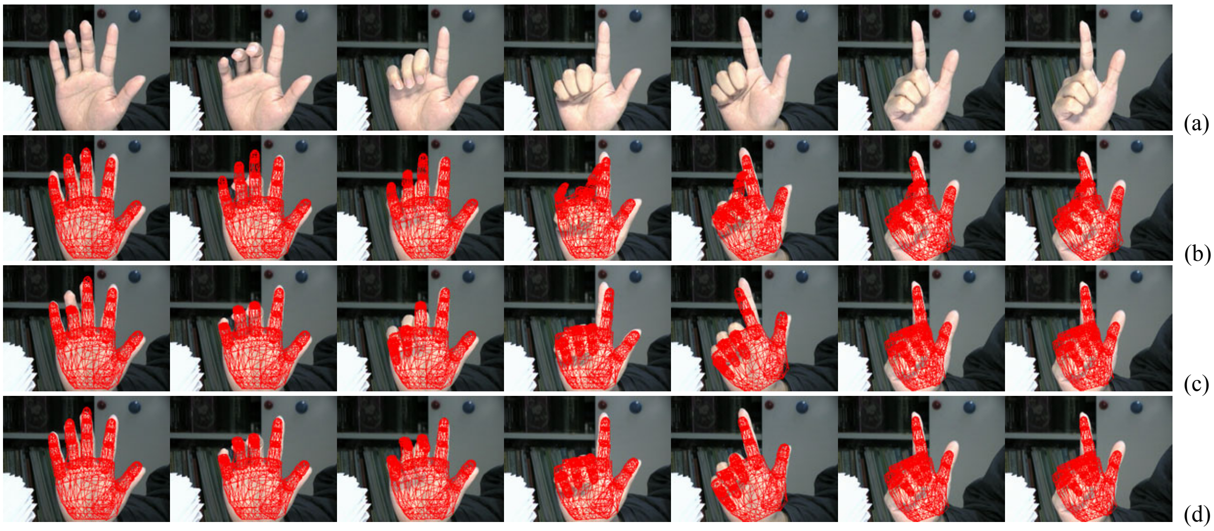


Figure 3. Global hand motion with three-finger articulation in 14 DOFs. (a) Input sequence. (b) Tracking results using classical particle filtering. (c) Tracking results using appearance information only. (d) Tracking results using our hybrid approach.

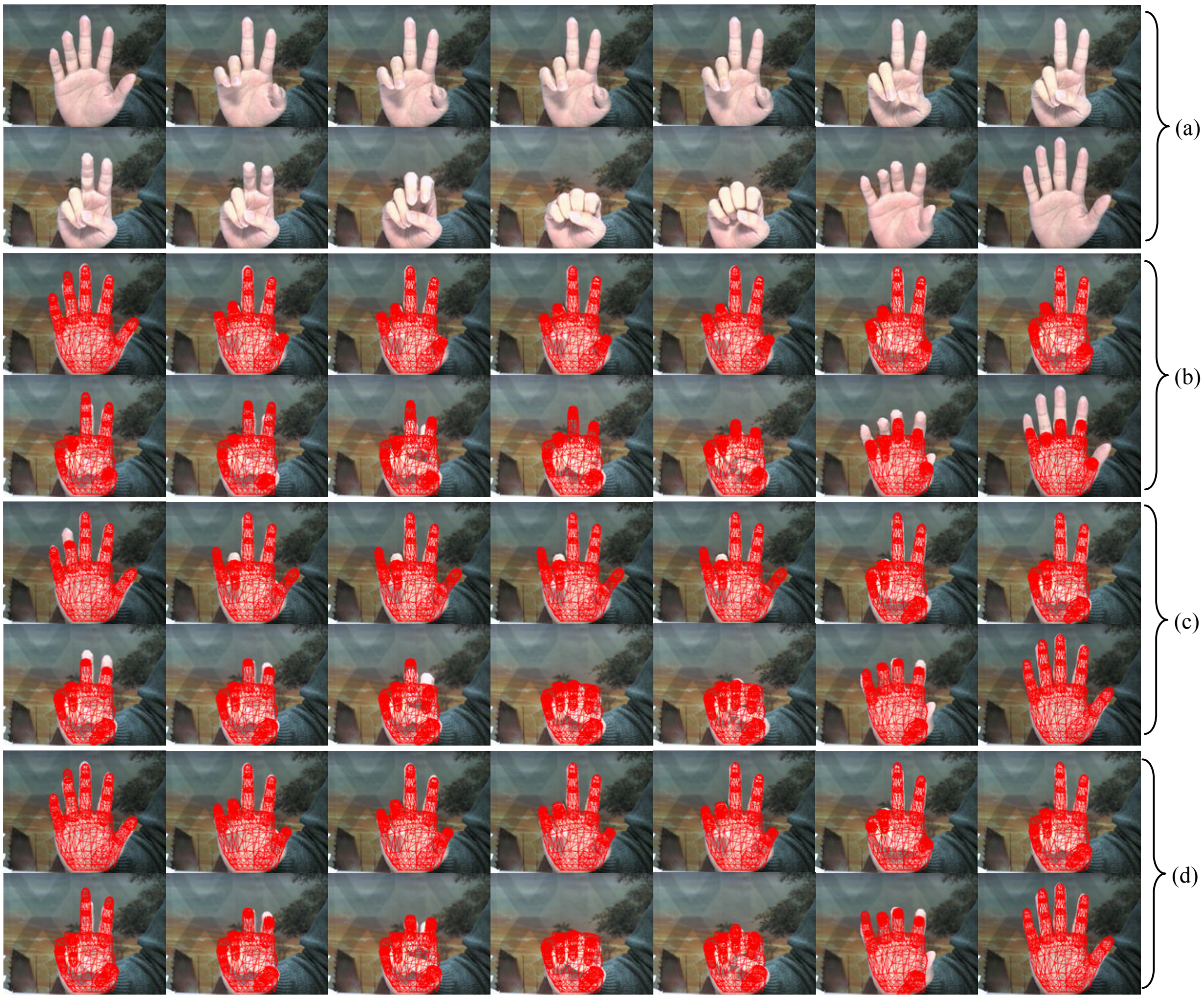


Figure 4. Five-finger tracking with 20 DOFs. (a) Input sequence. (b) Tracking results using classical particle filtering. (c) Tracking results using appearance information only. (d) Tracking results using our hybrid approach.

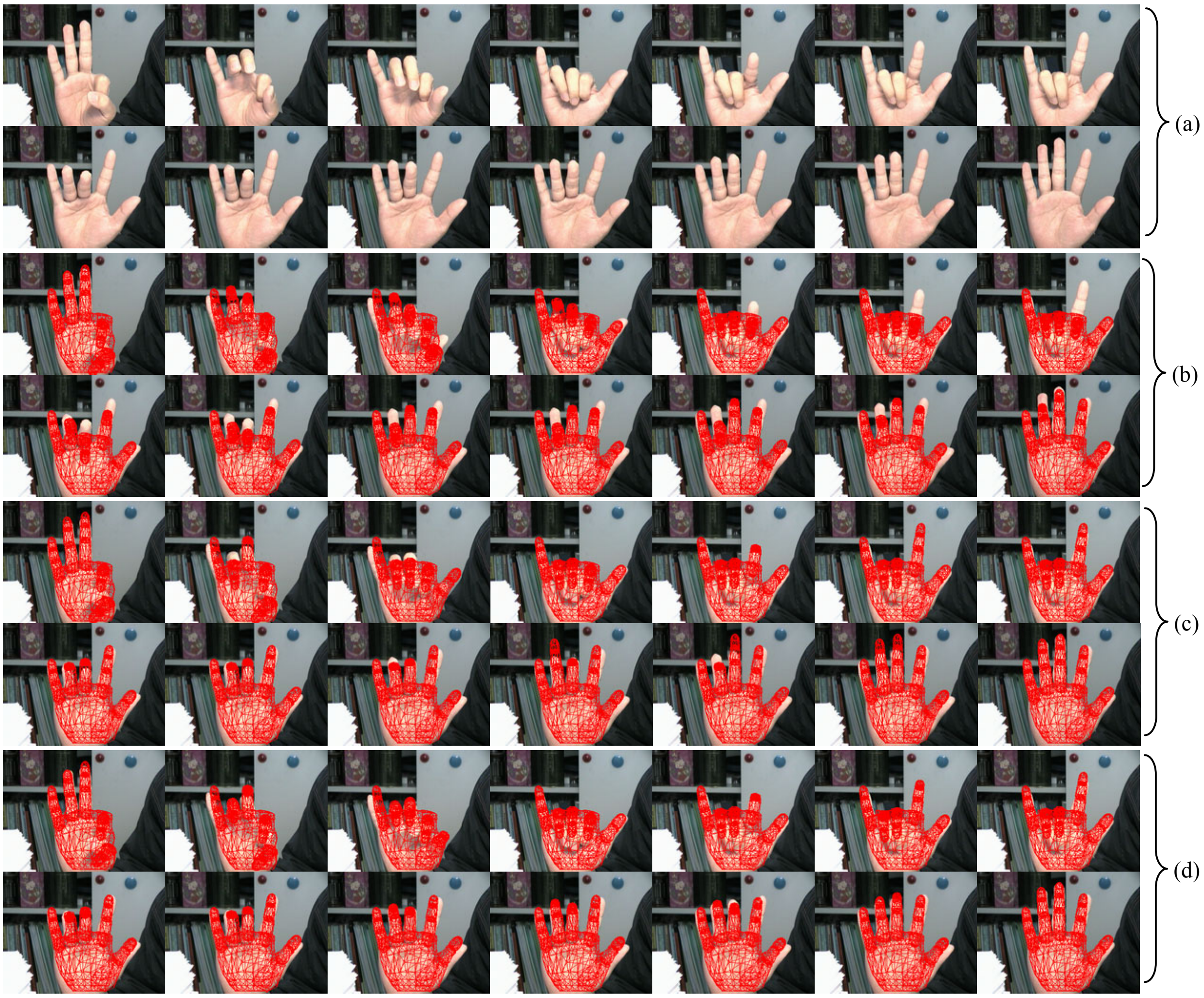


Figure 5. Four-finger tracking with 16 DOFs. (a) Input sequence. (b) Tracking results using classical particle filtering. (c) Tracking results using appearance information only. (d) Tracking results using our hybrid approach.

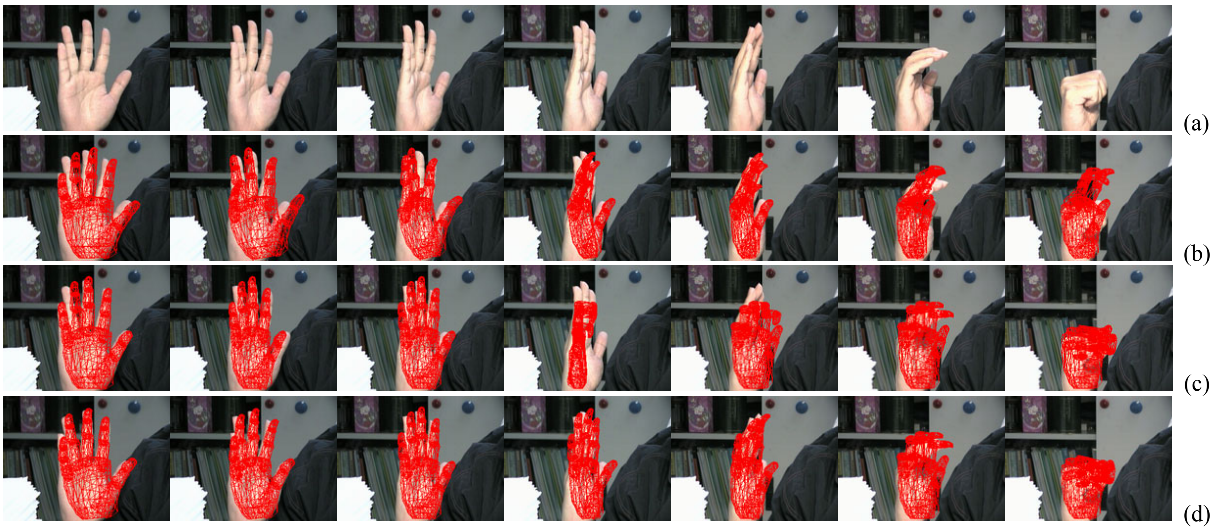


Figure 6. Global hand motion with five-finger articulation in 22 DOFs. (a) Input sequence. (b) Tracking results using classical particle filtering. (c) Tracking results using appearance information only. (d) Tracking results using our hybrid approach.