Appearance-Guided Particle Filtering for Articulated Hand Tracking

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

We propose a model-based tracking method, called appearance-guided particle filtering (AGPF), which integrates both sequential motion transition information and appearance information. A probability propagation model is derived from a Bayesian formulation for this framework, and a sequential Monte Carlo method is introduced for its realization. We apply the proposed method to articulated hand tracking, and show that it performs better than methods that only use either sequential motion transition information or only use appearance information.

1. Introduction

High degree-of-freedom (DOF) tracking, such as articulated hand tracking in arbitrary situations, is a challenging task. In general, there are two approaches suggested for high DOF tracking. One is the appearance-based approach [2][9][13][14] that estimates articulated motion states directly from images by learning the mapping from an image feature space to the object state space [9]. The other is the model-based approach [3][7][12][15][16] that estimates articulated motion states by projecting a 3D model on to the image space and then compares the projections with the observations. One advantage of the former is that observations from arbitrary viewpoints can be processed. However, large and dense reference images should be collected in advance to get an accurate estimation. Also, effective learning or retrieval in a large image set is very demanding. The latter approach can provide an accurate estimation when a 3D model is well initialized, but searching in a high-dimensional space is very complex.

In the model-based approach, the motion state is recovered from the 3D configuration with the maximal similarity. This problem has been formulated as an optimization problem [8][15], and can also be treated in a probabilistic framework as the state estimation of a dynamic system. Since closed-form solutions of a highly non-linear dynamic system are intractable, sequential Monte Carlo methods such as particle filtering were introduced to solve this problem. In the past, Isard and Blake [5] introduced the concept of particle filtering to visual tracking, named CONDENSATION. Rui and Chen [10] integrated unscented Kalman filter and particle filtering to generate a better proposal distribution. Particle filtering has been widely used in articulated hand tracking [3][7][16] recently. Wu *et al.* [16] suggested a method to represent the motion state in low-dimensional space by a set of linear manifolds constructed from base configurations and used particle filtering to track. Bray *et al*. [3] integrated the stochastic meta-descent optimization into particle filtering to find good particles for tracking, while Lin *et al*. [7] proposed a stochastic simplex search algorithm by combining the Nelder-Mead algorithm with particle filtering in a feasible space.

However, most of the particle filtering-based high DOF hand-tracking methods only use visual information from previous time steps. Although applying state estimators to a dynamic system has been shown to be effective for visual tracking, it has certain limitations. First, only initial states are employed; thus, the tracking process may 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. To overcome these difficulties, we study the state estimation of a dynamic system under the assumption that there are some known *attractors*, in addition to the initial state, in the state space. In this paper, an attractor is referred to as a state space vector whose observation is known. For a visual tracking problem, attractors are some reference images of the objects with known motion states, and serve as prior knowledge to guide the tracking in a high-dimensional space.

2. Statistical Model and Its Derivations

2.1. Observation Model for Hand Tracking

Let the state parameter vector of a target at time t be denoted as $\boldsymbol{x}_t$, and its observation as $\boldsymbol{z}_t$. The history of observations from time 1 to t is denoted as $\boldsymbol{Z}_t = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t\}$. A generic 3D hand model that has 22 DOFs is used for

* The corresponding author.

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

hand tracking in our work, where each finger has 4 DOFs and the palm has 2 DOFs of rotation. In our current implementation, the likelihood $p(\boldsymbol{z}_t|\ \boldsymbol{x}_t)$ is measured by using hand silhouettes. A hand area $\boldsymbol{H}_R$ is rendered from the 3D hand model under the state $\boldsymbol{x}_t$. For each input hand area $\boldsymbol{H}_I$, we define the likelihood by calculating the difference area between $\boldsymbol{H}_I$ and $\boldsymbol{H}_R$. In practice, two forces are formulated to minimize the difference area. One is the *shrink force*, $\boldsymbol{F}_{shrink}$, which has a tendency to shrink fingers of the 3D model to minimize the area, $E_{shrink} = \boldsymbol{H}_R - (\boldsymbol{H}_I \cap \boldsymbol{H}_R)$. The other force is the *stretch force*, $\boldsymbol{F}_{stretch}$, which has a tendency to expand fingers of the 3D hand model to minimize the difference area, $E_{stretch} = \boldsymbol{H}_I - (\boldsymbol{H}_I \cap \boldsymbol{H}_R)$. The likelihood is defined by combining these two forces. Then, the likelihood is defined as

$$p(\boldsymbol{z}_t|\ \boldsymbol{x}_t)$$

$$\propto \exp(-[w_1 \cdot Area(E_{shrink}) + w_2 \cdot Area(E_{stretch})]^2 / 2\sigma^2), \qquad (1)$$

where $Area(E)$ denotes the area of E, σ is a variance constant, and w_1 and w_2 are two weights.

Note that when $w_1 = w_2 = 0.5$, $p(\boldsymbol{z}_t|\ \boldsymbol{x}_t) \propto$

$$\exp\left(-Area[(\boldsymbol{H}_I \cup \boldsymbol{H}_R) - (\boldsymbol{H}_I \cap \boldsymbol{H}_R)]^2 / 8\sigma^2\right),$$

which is the total area of the non-overlapping regions.

2.2. Bayesian Formulation of AGPF

Assume that there are n attractors, $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_n$, in the state space. These attractors are pre-collected training samples whose observations are known and 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 $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_n$. With such prior information, the problem we focus on is as follows:

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

To achieve this, we investigate the probability:

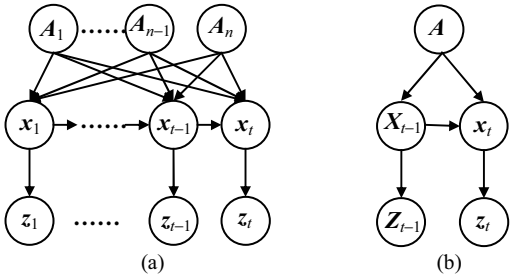
$$p(\boldsymbol{x}_t|\ \boldsymbol{Z}_t,\ \boldsymbol{A}). \qquad (2)$$

Suppose that

(i) The history of observations, $\boldsymbol{Z}_t$, is conditionally independent of the attractors, $\boldsymbol{A}$, given $\boldsymbol{X}_t$.

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

The dynamic Bayesian network (BN) structure considered in our work is shown in Fig. 1(a). Note that if the nodes and links in associations with $\boldsymbol{A}$ are all removed, it degenerates to a BN structure used for the classical particle filtering, where simply a first-order Markov chain is concerned; thus, the states are only influenced by previous time steps.



(a) (b)

Figure 1. (a) The dynamic BN structure of AGPF. (b)A concise representation.

By condensing $\{\boldsymbol{A}_1,\ldots,\boldsymbol{A}_n\}$, $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{t-1}\}$, and $\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{t-1}\}$ to the super nodes $\boldsymbol{A}$, $\boldsymbol{X}_{t-1}$, and $\boldsymbol{Z}_{t-1}$ in the BN, respectively, (as shown in Fig. 1(b)), it is easy to observe from the D-separation property [11] that there are some other conditional independencies inherent in the BN as shown below:

(iii) The observation at time t, $\boldsymbol{z}_t$, is conditionally independent of $\boldsymbol{Z}_{t-1}$ and $\boldsymbol{X}_{t-1}$, given the state $\boldsymbol{x}_t$.

(iv) The state at time t, $\boldsymbol{x}_t$, is conditionally independent of $\boldsymbol{Z}_{t-1}$, given the previous states $\boldsymbol{X}_{t-1}$.

From (i) to (iv), equation (2) can be resolved as

$$p(\boldsymbol{x}_t|\ \boldsymbol{Z}_t, \boldsymbol{A}) = \int_{\boldsymbol{x}_1 \ldots \boldsymbol{x}_{t-1}} p(\boldsymbol{X}_t|\ \boldsymbol{Z}_t, \boldsymbol{A}) \propto \int_{\boldsymbol{x}_1 \ldots \boldsymbol{x}_{t-1}} p(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{A})$$

$$= \int_{\boldsymbol{x}_1 \ldots \boldsymbol{x}_{t-1}} p(\boldsymbol{A}|\ \boldsymbol{X}_t, \boldsymbol{Z}_t) \cdot p(\boldsymbol{X}_t, \boldsymbol{Z}_t)$$

$$= \int_{\boldsymbol{x}_1 \ldots \boldsymbol{x}_{t-1}} p(\boldsymbol{A}|\ \boldsymbol{X}_t) \cdot p(\boldsymbol{X}_t, \boldsymbol{Z}_t) \qquad (3)$$

$$= \int_{\boldsymbol{x}_1 \ldots \boldsymbol{x}_{t-1}} p(\boldsymbol{A}|\ \boldsymbol{X}_t) \cdot p(\boldsymbol{z}_t|\ \boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t|\ \boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1}) \cdot p(\boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1})$$

$$= \int_{\boldsymbol{x}_1 \ldots \boldsymbol{x}_{t-1}} p(\boldsymbol{A}|\ \boldsymbol{X}_t) \cdot p(\boldsymbol{z}_t|\ \boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t|\ \boldsymbol{X}_{t-1}) \cdot p(\boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1}). \qquad (4)$$

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

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

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

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

Note that from (3),

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

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

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

Equation (6) relates the posterior probabilities $p(\boldsymbol{x}_t|\,\boldsymbol{Z}_t, \boldsymbol{A})$ to $p(\boldsymbol{x}_{t-1}|\,\boldsymbol{Z}_{t-1}, \boldsymbol{A})$ recursively, which shows how the posterior probabilities propagate given the prior probabilities $\boldsymbol{z}_t$ and $\boldsymbol{A}$. Like the original particle filtering, the MAP estimation of $\boldsymbol{x}_t$ can be iteratively obtained from previous time steps. A major distinction is that $\boldsymbol{x}_t$ is further affected by the attractors in $\boldsymbol{A}$ which contain prior appearance information.

3. Realization of AGPF

The Bayesian formulation of classical particle filtering is expressed as

$$p(\boldsymbol{x}_t|\boldsymbol{Z}_t) \propto p(\boldsymbol{z}_t|\boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t|\boldsymbol{Z}_{t-1}), \quad (7)$$

To precisely compute the posterior probability $p(\boldsymbol{x}_t|\boldsymbol{Z}_t)$ for obtaining a Bayesian optimal solution is computationally infeasible for nonlinear/non-Gaussian systems. In particle filtering, sequential Monte Carlo methods using importance sampling or re-sampling have been adopted to realize the computations [1][6][10]. The use of importance sampling has been shown to be a powerful methodology for sequential signal processing, since it can cope with difficult nonlinear and/or non-Gaussian problems.

However, tracking with particle filtering that depends only on the previous time step and the current frame suffers from undesirable drifting effects, particularly for long image sequences and/or high DOF moving objects. Unlike classical particle filtering in which tracking employs only sequential state probability propagation information, we have shown (in Section 2.2) how to obtain Bayesian optimal solutions when further prior knowledge of object appearances is imposed. A suitable importance sampling strategy for finding the solutions, referred to as AGPF, is derived below.

To realize (6), like classical particle filtering, we represent the posterior distribution $p(\boldsymbol{x}_{t-1}|\,\boldsymbol{Z}_{t-1}, \boldsymbol{A})$ by a set of weighted samples $\{(\boldsymbol{s}_{t-1}^{i}, \pi_{t-1}^{i}), i = 1,\ldots, N\}$ at time-step $t-1$, where $\boldsymbol{s}_{t-1}^{i}$ is the sample drawn from $p(\boldsymbol{x}_{t-1}|\,\boldsymbol{Z}_{t-1}, \boldsymbol{A})$ and π_{t-1}^{i} is the associated weight. The weights are chosen using the principle of importance sampling [4] with $\pi_t \propto p(\boldsymbol{X}_t|\,\boldsymbol{Z}_t, \boldsymbol{A})/q(\boldsymbol{X}_t|\,\boldsymbol{Z}_t, \boldsymbol{A})$, where $q(\boldsymbol{X}_t|\,\boldsymbol{Z}_t, \boldsymbol{A})$ is an importance distribution (or proposal function) from which it is easier to draw samples than from the probability density $p(\boldsymbol{X}_t|\,\boldsymbol{Z}_t, \boldsymbol{A})$.

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 $q(\boldsymbol{X}_t|\,\boldsymbol{Z}_t, \boldsymbol{A}) = q(\boldsymbol{x}_t|\,\boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}) \cdot q(\boldsymbol{X}_{t-1}|\,\boldsymbol{Z}_{t-1}, \boldsymbol{A})$, where $\boldsymbol{s}_t^{i}$ can be generated from $q(\boldsymbol{x}_t|\,\boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A})$ and $\boldsymbol{S}_{t-1}^{i} = \{\boldsymbol{s}_1^{i},\ldots, \boldsymbol{s}_{t-1}^{i}; i = 1,\ldots, N\}$ can be generated from $q(\boldsymbol{X}_{t-1}|\,\boldsymbol{Z}_{t-1}, \boldsymbol{A})$. Consider that

$$\pi_t \propto p(\boldsymbol{X}_t|\,\boldsymbol{Z}_t, \boldsymbol{A})/q(\boldsymbol{X}_t|\,\boldsymbol{Z}_t, \boldsymbol{A}) \propto p(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{A})/q(\boldsymbol{X}_t|\,\boldsymbol{Z}_t, \boldsymbol{A}). \quad (8)$$

In addition, according to (3) and (4), the term $p(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{A})$ can be rewritten as

$$\begin{aligned} p(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{A}) &= p(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A}) \cdot p(\boldsymbol{A}|\,\boldsymbol{X}_{t-1}) \cdot p(\boldsymbol{z}_t|\,\boldsymbol{x}_t) \cdot p(\boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1}) \\ &\propto p(\boldsymbol{z}_t|\,\boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A}) \cdot p(\boldsymbol{X}_{t-1}|\,\boldsymbol{Z}_{t-1}, \boldsymbol{A}). \end{aligned} \quad (9)$$

By substituting (9) into (8), we have

$$\begin{aligned} \pi_t &\propto \pi_{t-1} \cdot [p(\boldsymbol{z}_t|\,\boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A})]/[q(\boldsymbol{x}_t|\,\boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A})] \quad (10) \\ &= \pi_{t-1} \cdot [p(\boldsymbol{z}_t, \boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A})]/[q(\boldsymbol{x}_t|\,\boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A})]. \end{aligned}$$

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

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

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

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

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

$$\begin{aligned} q(\boldsymbol{x}_t|\,\boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}) &= p(\boldsymbol{z}_t|\,\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{A}) \cdot p(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A})/p(\boldsymbol{z}_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A}) \\ &= p(\boldsymbol{z}_t|\,\boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A})/p(\boldsymbol{z}_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A}). \end{aligned} \quad (11)$$

Substituting (11) into (10), the optimal weight is

$$\pi_t \propto \pi_{t-1} \cdot p(\boldsymbol{z}_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A}) = \pi_{t-1} \int_{\boldsymbol{x}'_t} p(\boldsymbol{z}_t|\,\boldsymbol{x}'_t) \cdot p(\boldsymbol{x}'_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A}). \quad (12)$$

However, the optimal importance weight (12) is difficult to evaluate since an integral is needed. In practice, we choose $q(\boldsymbol{x}_t|\,\boldsymbol{X}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{A}) = p(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}, \boldsymbol{A})$ 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 $\boldsymbol{A}$ is imposed. In this case, $\pi_t \propto \pi_{t-1} \cdot p(\boldsymbol{z}_t|\,\boldsymbol{x}_t)$, which is the same as that in classical particle filtering.

4. AGPF Filtering Distributions

We set the distribution of $p(\boldsymbol{x}_t|\ \boldsymbol{x}_{t-1},\ \boldsymbol{A})$ as a mixture of $p(\boldsymbol{x}_t|\ \boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_t|\ \boldsymbol{A}_i)$, where $i = 1, \ldots, n$.

$$p(\boldsymbol{x}_t|\ \boldsymbol{x}_{t-1},\ \boldsymbol{A}) = \alpha_0 \cdot p(\boldsymbol{x}_t|\ \boldsymbol{x}_{t-1}) + \Sigma_{i=1,\ldots,n} \alpha_i \cdot p(\boldsymbol{x}_t|\ \boldsymbol{A}_i), \quad (13)$$

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

In (13), the probability $p(\boldsymbol{x}_t|\ \boldsymbol{x}_{t-1})$ is the probability caused by state transition, and $p(\boldsymbol{x}_t|\ \boldsymbol{A})$ is the probability caused by appearance similarity. By substituting (13) into (6), we have

$$p(\boldsymbol{x}_t|\ \boldsymbol{Z}_t, \boldsymbol{A}) \propto p(\boldsymbol{z}_t|\ \boldsymbol{x}_t) \cdot [\Sigma_{i=1,\ldots,n} \alpha_i \cdot p(\boldsymbol{x}_t|\ \boldsymbol{A}_i) + \int_{\boldsymbol{x}_{t-1}} \alpha_0 \cdot p(\boldsymbol{x}_t|\ \boldsymbol{x}_{t-1}) \cdot p(\boldsymbol{x}_{t-1}|\ \boldsymbol{Z}_{t-1}, \boldsymbol{A})]. \quad (14)$$

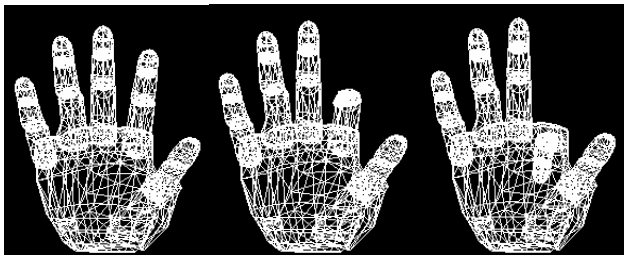
When α_0 is set to one, we obtain the classical particle filtering. In contrast, if α_0 is set to zero, our method degenerates to a pure appearance-based approach.

In AGPF, 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 $p(\boldsymbol{x}_t|\ \boldsymbol{A})$ in our approach, 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 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 (6), M_1 is set as $\alpha_0 \cdot N$ and M_2 is set as $(1-\alpha_0) \cdot N$. Hence, the sample set at time-step t $\{\boldsymbol{s}_t^i,\ i = 1,\ldots,\ N\}$ is $\{\hat{\boldsymbol{s}}_t^k\} \cup \{\boldsymbol{a}^j\}$. The weight π_t^i of each sample in $\{\boldsymbol{s}_t^i\}$ is calculated from the observation distribution $p(\boldsymbol{z}_t|\ \boldsymbol{x}_t = \boldsymbol{s}_t^i)$ by comparing the similarity between observation $\boldsymbol{z}_t$ and the state $\boldsymbol{x}_t$. Finally, the desired posterior distribution $p(\boldsymbol{x}_t|\ \boldsymbol{Z}_t, \boldsymbol{A})$ 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

$$\boldsymbol{x}_t^* = \boldsymbol{s}_t^*, \text{ where } \pi_t^* = \max(\pi_t^i). \quad (15)$$

5. AGPF-based Hand Tracking

A primary difficulty of articulated hand tracking is that the motion DOF is too high, resulting in too many possible appearances. Although there are some approaches that use a large quantity of pre-collected appearances [2][9], collecting all the appearances for comparison is still too hard to be feasible. On the other hand, by starting from the initial state in association with an articulated configuration, all the configurations (in association with their appearances) can be reached by gradually changing the articulated motion parameters. This is a significant reason why state-space methods (such as particle filtering) have been widely used in recent studies. However, since articulated configurations are estimated only from sequential updating information, state-space methods easily mis-track, especially when motions between two consecutive images are large. The AGPF method integrates both the motion transition model and appearance information, and thereby avoids both difficulties.



(a) (b) (c)

Figure 2. Three appearances of the index finger.

To construct the appearances for AGPF, only a limited number of the appearances are needed, since sequential motion-transition information is also available. In our work, three appearances for each finger and nine appearances for global hand motion are pre-collected by bending the finger and rotating the palm to different levels. An example of the appearances of the index finger is shown in Fig. 2.

In our framework, all the attractors affect the state to be estimated with certain probabilities. Nevertheless, to make our implementation more efficient, we use the most significant K attractors which have the largest probability values of $p(\boldsymbol{z}_t|\ \boldsymbol{A}_i)$ instead. As the attractors far away from the current state usually have small probabilities that can be neglected, our method becomes more efficient by using this strategy.

6. Experimental Results

In our experiments, we compare the tracking results of the AGPF method with those obtained by either using classical particle filtering, or by using appearance information only. Image sequences with a large range of motions are captured. Twenty particles are used for each experiment, i.e., $N = 20$. The probabilities $p(\boldsymbol{x}_t|\ \boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_t|\ \boldsymbol{A}_i)$ are modeled by Gaussian distributions,

$$p(\boldsymbol{x}_t|\ \boldsymbol{x}_{t-1}) \sim \mathcal{N}(\boldsymbol{x}_{t-1}, \boldsymbol{\Sigma}_1) \text{ and } p(\boldsymbol{x}_t|\ \boldsymbol{A}_i) \sim \mathcal{N}(\boldsymbol{A}_i, \boldsymbol{\Sigma}_2),$$

respectively, where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are diagonal covariance matrices.

In the first experiment (Fig. 3), K is set as one and an 8-DOF model obtained by bending two fingers is used. There are $3^2 = 9$ attractors, where 3 attractors are

generated for each finger. Fig. 3(a) shows part of the input images. The tracking results of classical particle filtering, the pure appearance-based approach (by setting α_0 as zero), and the AGPF method are shown in Figs. 3(b), 3(c), and 3(d), respectively. From Fig. 3(b), one can see that classical particle filtering fails to track the third, fifth, and sixth images. In addition, tracking also is not very accurate when using pure appearance information, as shown in Fig. 3(c), particular when no pre-collected images match the current motion state (e.g., the second image in Fig. 3(c)). By contrast, the proposed AGPF successfully recovers these motions, as shown in Fig. 3(d).

In the second experiment (Fig. 4), a 14-DOF model and 9×3^3 = 243 attractors (containing 3 attractors for each finger and 9 attractors for global hand rotation) are used, and K is set to one. As with the first experiment, the tracking results from using classical particle filtering or pure appearance information are not very accurate in this experiment. Fig. 4 (b) shows that classical particle filtering fails to track all of the images, except the first one. Furthermore, the fourth image in Fig. 4(c) shows that tracking also is not very accurate when using pure appearance information. However, our proposed method provides better results, as shown in Fig. 4(d).

Some other experiments were performed with different DOFs and different numbers of attractors. Fig. 5 shows the results of tracking four fingers with 16 DOFs. In this experiment, 81 attractors are used. Another four-finger tracking with 81 attractors is shown in Fig. 6. Finally, Fig. 7 shows the tracking results of global hand motion with five-finger articulation, where the DOF is 22 with 2,187 attractors. The K value sets for the above three experiments are one, three, and one, respectively.

To demonstrate how the tracking trajectory is influenced by the attractors, we use the first experiment as an example. Fig. 8 shows the tracking trajectories of classical particle filtering and AGPF, where principal component analysis (PCA) is used to reduce the high dimensional representation of samples to a 2D space for visualization. In Fig. 8, the blue points and line respectively represent the samples and trajectory generated by classical particle filtering, and the red points and line respectively represent those generated by the AGPF method. The point sign, •, represents the samples generated from the motion transition model and the star sign, ∗, represents the samples generated by attractors. In addition, the attractor with the maximum probability $p(\boldsymbol{z}_t|\, \boldsymbol{A}_i)$ is shown as a green circle. From this figure, it is obvious that attractors can guide the motion trajectory toward a more accurate tracking result.

One can see from the experiments that, with a limited number of pre-collected appearances, our approach significantly refines the performance of the approach of classical particle filtering in which only a motion transition model is used, and also performs better than the approach when only appearances are used.

7. Conclusions

In this paper, we give a statistically optimized framework for tracking that considers both pre-collected appearance and on-line motion transition information. A probability propagation model that integrates both types of information is derived. With the pre-collected appearances, the proposed method can still be performed efficiently since a sequential Monte Carlo method is provided. By applying prior appearance information, our method can recover rapid motions that are difficult to solve by using simple particle filtering. In addition, unlike pure appearance approaches, 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-selected in the appearance database. Promising tracking results are obtained by using the proposed method in our experiments.

Acknowledgment: This work was supported in part under Grants NSC 93-2752-E-002-007-PAE, NSC 93-2213-E-001-022, 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. Isard and A. Blake, "ICONDENSATION: unifying low-level and high-level tracking in a stochastic framework," *Proc. European Conf. Computer Vision*, pp.893-908, 1998.

[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] S. Lu, D. Metaxas, D. Samaras, and J. Oliensis, "Using multiple cues for hand tracking and model refinement," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 443-450, Vol. 2, 2003.

[9] 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.

[10] Y. Rui and Y. Chen, "Better proposal distributions: object tracking using unscented particle filter," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 786-793, Vol. 2, 2001.

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

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

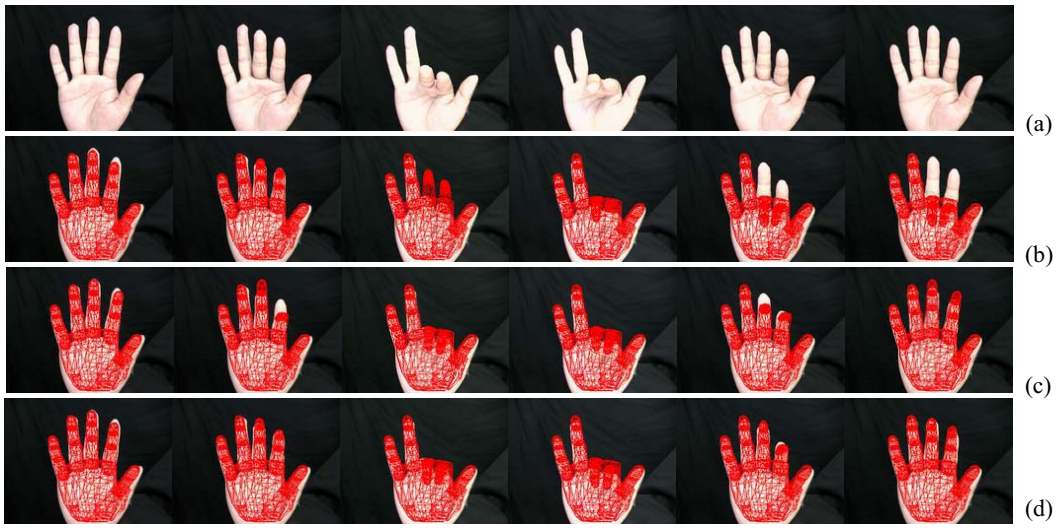


Figure 3. Two-finger tracking with 8 DOFs. (a) Input sequence. (b) Tracking results by using classical particle filtering. (c) Tracking results by using appearance information only. (d) Tracking results by using the AGPF method.

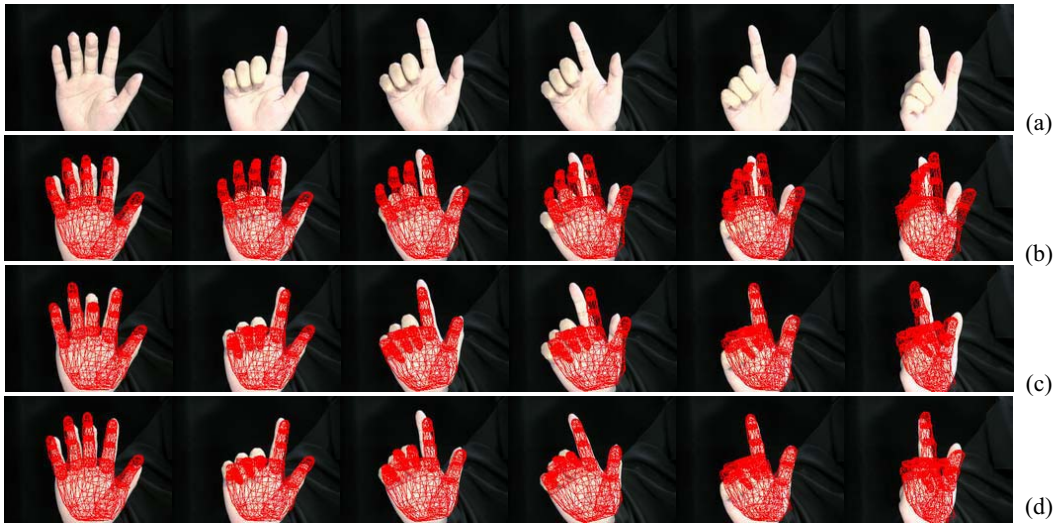


Figure 4. Global hand motion with three-finger articulation in 14 DOFs. (a) Input sequence. (b) Tracking results by using classical particle filtering. (c) Tracking results by using appearance information only. (d) Tracking results by using the AGPF method.

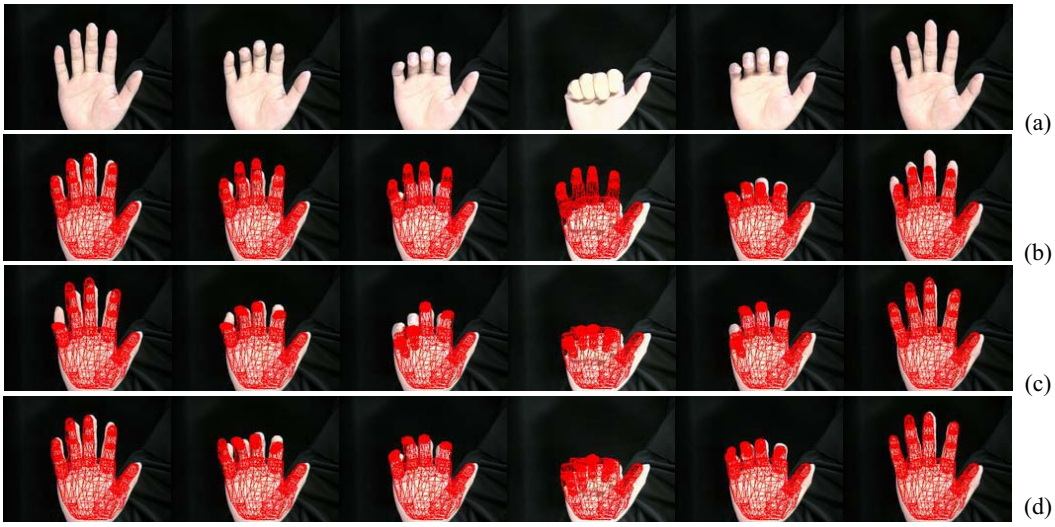


Figure 5. Four-finger tracking with 16 DOFs. (a) Input sequence. (b) Tracking results by using classical particle filtering. (c) Tracking results by using appearance information only. (d) Tracking results by using the AGPF method.

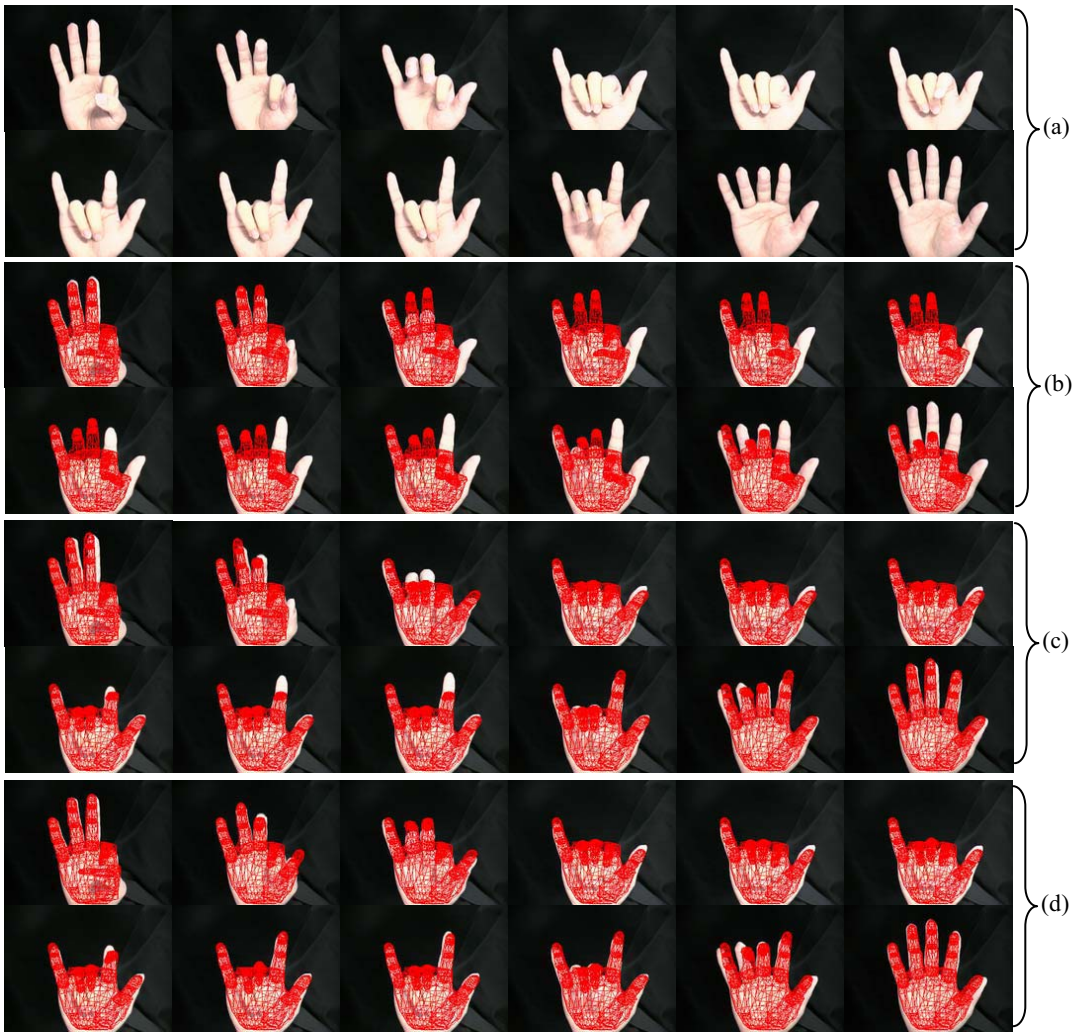


Figure 6. Four-finger tracking with 16 DOFs. (a) Input sequence. (b) Tracking results by using classical particle filtering. (c) Tracking results by using appearance information only. (d) Tracking results by using the AGPF method.

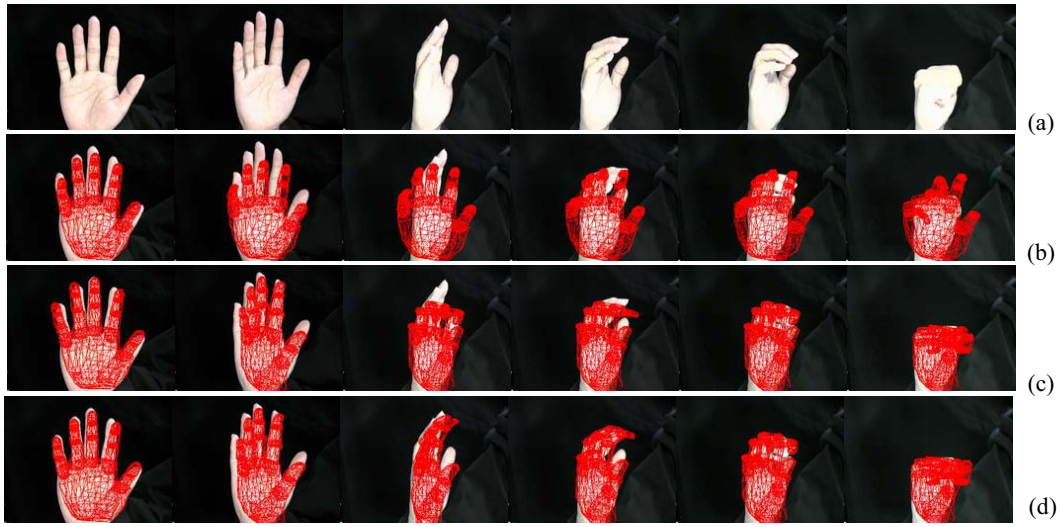


Figure 7. Global hand motion with five-finger articulation in 22 DOFs. (a) Input sequence. (b) Tracking results by using classical particle filtering. (c) Tracking results by using appearance information only. (d) Tracking results by using the AGPF method.

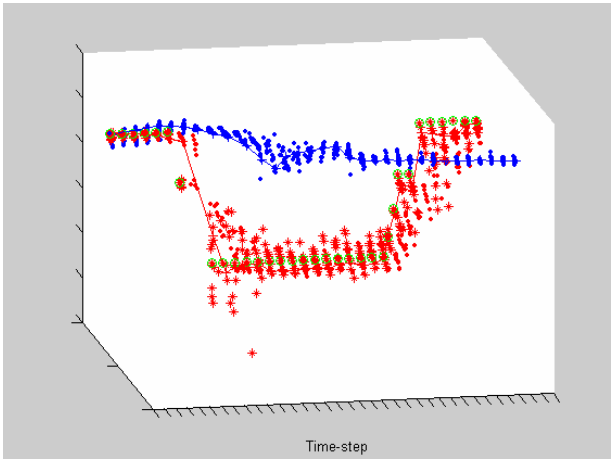


Figure 8. Tacking trajectories reduced to a 2D PCA space of classical particle filtering and the AGPF of the first experiment (Figure 3). The blue points and line respectively represent the samples and trajectory generated by classical particle filtering, and the red points and line respectively represent those generated by the AGPF method. The point sign, •, represents the samples generated from the motion transition model and the star sign, ∗, represents the samples generated by attractors. In addition, the attractor with the maximum probability $p(\boldsymbol{z}_t|\boldsymbol{A}_i)$ is shown as a green circle. From this figure, it is obvious that attractors can guide the motion trajectory toward a more accurate tracking result.