Visual Tracking in High-Dimensional State Space by Appearance-Guided Particle Filtering

Wen-Yan Chang, Chu-Song Chen, and Yong-Dian Jian

***Abstract*—In this paper, we propose a new approach, appearance-guided particle filtering (AGPF), for high degree-of-freedom visual tracking from an image sequence. This method adopts some known attractors in the state space and integrates both appearance and motion-transition information for visual tracking. A probability propagation model based on these two types of information is derived from a Bayesian formulation, and a particle filtering framework is developed to realize it. Experimental results demonstrate that the proposed method is effective for high degree-of-freedom visual tracking problems, such as articulated hand tracking and lip-contour tracking.**

***Index Terms*—Appearance-guided particle filtering (AGPF), articulated hand tracking, lip-contour tracking, particle filtering, sequential Monte Carlo method, visual tracking.**

I. Introduction

TRACKING in high-dimensional space is a challenging problem. Unlike the 2-D tracking, which focuses on locating the target in images, high degree-of-freedom (DOF) tracking further involves highly complex searching or matching. In recent years, high DOF tracking has been addressed on the topics of articulated hand and body gesture tracking in arbitrary situations. In general, two types of approach are used to solve this problem: *appearance-based* and *dynamic model-driven* methods.

In this paper, we mainly focus on articulated hand tracking; however, it can be demonstrated that our method is also applicable to other high DOF visual tracking problems. Appearance-based approaches employ mainly static appearance information about an object of interest. By collecting a set of images of distinct poses of the target object, articulated motion states can be estimated directly from the images by learning the mapping from an image feature space to the object state space [2], [3], [29], [33], [36]. In line with this research track, Rosales *et al.* [29] proposed the specialized mappings architecture (SMA), a state recovery method that learns the mapping between image features and their corresponding states for 3-D hand posture estimation. In [36], an image of an object is represented as a graph on which the nodes are labeled with a local image description and the edges are labeled with a distance vector. Elastic graph matching (EGM) with multiple features is then used to identify a proper posture. Athitsos and Sclaroff [2] formulated high DOF tracking as an image database indexing problem, and used a hierarchical retrieval method to find a proper state from a database containing images with simple backgrounds. To apply this concept to a cluttered environment and improve its performance, Euclidean embedding and probabilistic line matching methods are suggested in [3]. Furthermore, image matching by significant point groupings can also be used for appearance retrieval, as proposed in [7]. Stenger *et al.* [33] introduced a tree-based representation with a Bayesian filtering search technique to speed up tracking, while Shakhnarovich *et al.* [32] proposed a hashing-based approach for efficient searching of appearances.

In appearance-based approaches, precollected appearances can be treated as discrete samples in an object state space. However, a major difficulty with such approaches is that a large number of samples of appearances are required for high-DOF tracking in nearly arbitrary situations, which is infeasible in practice. On the other hand, dynamic model-driven approaches focus on dynamic and continuous information. In this type of approach, a dynamic system is formulated for visual tracking, and state-estimation techniques are suggested to solve it. Tracking in high DOF has been formulated as an optimization problem [26], [38]. In [38], Wu and Huang suggested a divide-and-conquer framework for tracking hand motion by dividing it into global motion and local finger motion. Lu *et al.* [26] injected multiple cues, including edges, optical flow, and shading information, into a deformable model to capture articulated hand motion in a simple environment. However, temporal coherence of sequential motions is not considered in these optimization methods.

For temporal information to be applied effectively, a dynamic system is formulated for state estimation in high-dimensional space so that estimation can be performed sequentially and optimally based on the system's dynamics. To this end, Isard and Blake [17] introduced particle filtering for visual tracking of a dynamic system based on sequential Monte Carlo estimation. As particle filtering can cope with difficult nonlinear/non-Gaussian problems, the methodology has been widely used for dynamic model-driven articulated hand tracking [6], [25], [34], [39] in recent years. Wu *et al.* [39] proposed a method that represents the motion state in low-dimensional space by a set of linear manifolds constructed from base configurations, and used particle filtering to track hand

Manuscript received June 18, 2006; revised March 13, 2008. This work was supported in part by Grants NSC 96-2752-E-002-007-PAE and 96-EC-17-A-02-S1-032. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Onur G. Guleryuz.

W.-Y. Chang is with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C., and also with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. (e-mail: wychang@iis.sinica.edu.tw).

C.-S. Chen is with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C., and also with the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan, R.O.C. (e-mail: song@iis.sinica.edu.tw).

Y.-D. Jian is with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C. (e-mail: ydjian@iis.sinica.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2008.924283

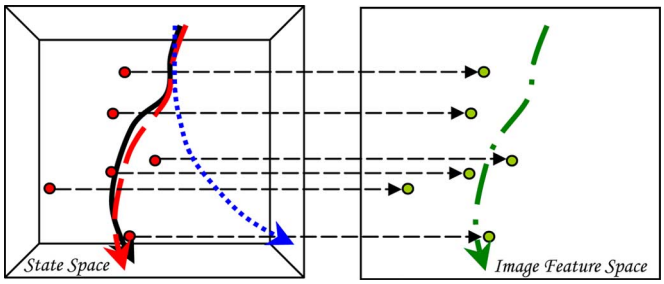


Fig. 1. Concept of attractors: Attractors in the state space are represented as red points and their corresponding appearances in the image feature space are represented as green points. The black solid curve represents the ground-truth, the green dash-dotted curve in the image feature space represents an observation sequence, and the blue dotted curve illustrates that tracking may easily drift when the posterior is evaluated solely on the observations in a motion-transition model. In contrast, with attractors, appropriate guidance can be given for tracking, as shown by the red dashed curve.

motions. Bray *et al.* [6] integrated stochastic meta-descent optimization into particle filtering to find good particles for tracking, while Lin *et al.* [25] proposed a stochastic simplex search algorithm that combines the Nelder–Mead algorithm with particle filtering in a feasible space. Sudderth *et al.* [34] used multiple independent trackers for each hand articulation, and applied Nonparametric Belief Propagation (NBP) to adjust particles iteratively for obtaining better estimations. In addition to particle filtering, Bors and Pitas [5] suggested a Bayesian approach for multiobject tracking and prediction in an image sequence. Based on the median radial basis function network, both the segmentation and the optical flow of moving objects can be estimated with their method.

Although applying state estimators to a dynamic system has been proven effective for visual tracking, most dynamic model-driven high-DOF tracking methods only use visual information from previous time steps. Some limitations thus arise. First, as tracking is only initiated from a single state whose appearance is known, the dynamic transition information from previous time steps plays a decisive role during tracking, and the tracking process may easily get trapped in local minima. Second, it is difficult to apply known object appearance information to boost the tracking performance, even when such information is easy to acquire.

To resolve these limitations, we study the state estimation problem of a dynamic system under the assumption that, in addition to the initial state, there are some known *attractors* in the state space. In this paper, an attractor is defined as a state space vector whose observation is known. For a visual tracking problem, attractors can be treated as some motion states of the objects with known reference images. In other words, an attractor is a static state containing appearance information in the image feature space, and serves as prior knowledge to guide the tracking in a high-dimensional state space. The concept is illustrated in Fig. 1. Note that, unlike some studies based on grid-based filtering [33], [35] 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. The attractor formalism can be seen as a specific constraint on a high-dimensional particle filtering.

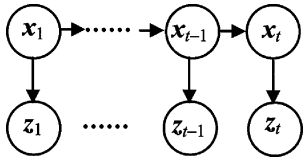


Fig. 2. Dynamic Bayesian network structure of particle filtering.

The advantages of considering attractors in a dynamic system include that they can regulate the configuration space for a high-DOF tracking problem, and can be precollected or pregenerated to boost the tracking performance of a dynamic model-driven approach. Given an observation sequence in the image feature space, we endeavor to find the *maximum a posteriori* (MAP) solution based on an attractor-regulated dynamic model. When the state of time in the dynamic model is represented as a first-order Markov chain, particle filtering is an efficient means of inferring an approximation of the optimal solution if no attractors are involved. However, it is not applicable formally to the case when attractors are applied. In this paper, we extend particle filtering to appearance-guided (or attractor-guided) particle filtering (AGPF), and derive a probability propagation framework to find its MAP solution.

The remainder of this paper is organized as follows. The AGPF framework is introduced in Section II. In Section III, we introduce the mixture-based AGPF and describe some filtering distributions used in our work. Sections IV–VI present the applications of AGPF to articulated hand tracking and lip-contour tracking, respectively. Section VII contains a discussion of AGPF. Then, in Section VIII, we present our conclusions.

II. Appearance-Guided Particle Filtering

We begin with a concise review of particle filtering and then introduce the AGPF framework.

A. Review of Particle Filtering

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 $\boldsymbol{Z}_t = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t\}$. The Bayesian formulation of particle filtering is expressed as

$$\begin{aligned} p(\boldsymbol{x}_t|\boldsymbol{Z}_t) &\propto p(\boldsymbol{z}_t|\boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t|\boldsymbol{Z}_{t-1}) \\ &= p(\boldsymbol{z}_t|\boldsymbol{x}_t) \cdot \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \cdot p(\boldsymbol{x}_{t-1}|\boldsymbol{Z}_{t-1}) d\boldsymbol{x}_{t-1} \end{aligned} \tag{1}$$

where a first-order Markov chain of the states is considered, and the likelihood $p(\boldsymbol{z}_t|\boldsymbol{x}_t)$ is the observation model. The observation, $\boldsymbol{z}_t$, is conditionally independent of the history of the observations from time 1 to $t-1$, $\boldsymbol{Z}_{t-1}$, given the state $\boldsymbol{x}_t$. The Bayesian network (BN) structure of particle filtering is shown in Fig. 2.

To compute the posterior probability, $p(\boldsymbol{x}_t|\boldsymbol{Z}_t)$, a closed-form solution with an integral over all possible state values in each iteration is formulated [1], [17]; however, it is computationally intractable. In particle filtering, sequential Monte Carlo methods using importance sampling or re-sampling have been adopted to realize the computations [1], [18]. The use of importance

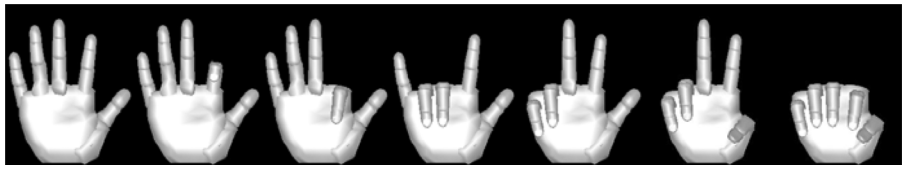


Fig. 3. Some attractors used in our work. The appearance of an attractor is obtained by rendering the associated 3-D hand model.

sampling has been shown to be a powerful strategy for sequential signal processing. With the dynamic/temporal propagation, particle filtering has been widely used for tracking applications [6], [25], [27], [39], [40]. Nevertheless, tracking with particle filtering that depends only on the previous time step and the current frame tends to cause undesirable drifting in high-DOF tracking, particularly for fast moving targets.

B. Probability Propagation of AGPF

Unlike particle filtering, in which the tracking only employs sequential state probability propagation information, we show how to obtain Bayesian optimal solutions when further prior knowledge of object appearances is introduced, and derive suitable importance sampling strategies for finding the solutions. 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 have been selected, their appearances can be generated by the observation model. For instance, in articulated hand tracking, appearances are generated from the projections of a generic 3-D hand model with distinct configuration parameters, examples of which are shown in Fig. 3. The generic model we currently use is textureless; thus, silhouettes of the rendered hand images serve as the main visual clues of appearances.

In this work, we assume that 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 estimation of state $\boldsymbol{x}_t$ is not only related to its previous state, $\boldsymbol{x}_{t-1}$, but also to the attractors $\boldsymbol{A} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n\}$. With the prior knowledge inherent in the appearances, we investigate the probability

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

Fig. 4(a) shows the BN structure considered in our framework. Note that if the nodes and links associated with $\boldsymbol{A}$ are removed, it degenerates to a BN structure for particle filtering, where only a first-order Markov chain is considered. By condensing $\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n\}$, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}\}$, and $\{z_1, \ldots, z_{t-1}\}$ as the super nodes $\boldsymbol{A}$, $\boldsymbol{X}_{t-1}$, and $\boldsymbol{Z}_{t-1}$ in the BN, respectively, as shown in Fig. 4(b), it can be derived from the D-separation property [31] that there are two properties.

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

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

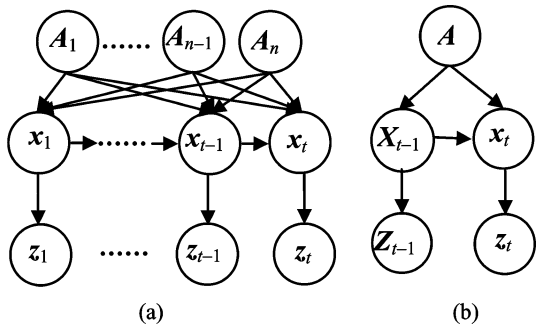


Fig. 4. (a) Dynamic Bayesian network structure of the AGPF. (b) Concise representation of the AGPF.

According to the above properties, (2) can be resolved as follows:

$$\begin{aligned} p(\boldsymbol{x}_t|\boldsymbol{Z}_t, \boldsymbol{A}) &\propto \int p(\boldsymbol{X}_t, \boldsymbol{Z}_t, \boldsymbol{A}) d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_{t-1} \\ &\propto \int p(\boldsymbol{z}_t|\boldsymbol{X}_t, \boldsymbol{Z}_{t-1}, \boldsymbol{A}) \cdot p(\boldsymbol{x}_t|\boldsymbol{X}_{t-1}, \boldsymbol{Z}_{t-1}, \boldsymbol{A}) \\ &\quad \cdot p(\boldsymbol{X}_{t-1}|\boldsymbol{Z}_{t-1}, \boldsymbol{A}) d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_{t-1} \\ &= \int p(\boldsymbol{z}_t|\boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{A}) \\ &\quad \cdot p(\boldsymbol{X}_{t-1}|\boldsymbol{Z}_{t-1}, \boldsymbol{A}) d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_{t-1} \\ &= p(\boldsymbol{z}_t|\boldsymbol{x}_t) \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{A}) \\ &\quad \cdot p(\boldsymbol{X}_{t-1}|\boldsymbol{Z}_{t-1}, \boldsymbol{A}) d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_{t-1} \\ &= p(\boldsymbol{z}_t|\boldsymbol{x}_t) \int p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{A}) \\ &\quad \cdot p(\boldsymbol{x}_{t-1}|\boldsymbol{Z}_{t-1}, \boldsymbol{A}) d\boldsymbol{x}_{t-1}. \end{aligned} \tag{3}$$

Equation (3) 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 $\boldsymbol{A}$. Likewise, the MAP estimation of $\boldsymbol{x}_t$ can be iteratively obtained from previous time steps. Compared to the formulation in (1), there is a major distinction between the original particle filtering and that of AGPF. In the MAP solution (3), 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.

C. Sequential Monte Carlo Framework of AGPF

The probability propagation solutions of (3) are still computationally infeasible, since the integral over all possible state values is too complex to evaluate. To realize (3), as in standard particle filtering, we represent the posterior distribution $p(\boldsymbol{x}_t|\boldsymbol{Z}_t, \boldsymbol{A})$ by a set of weighted samples $\left\{\left(\boldsymbol{s}_t^i, \pi_t^i\right), i = 1, \ldots, N\right\}$ at time step t, where $\boldsymbol{s}_t^i$ is the

sample and π_t^i is the associated weight. The weights are chosen by using the principle of importance sampling [14] 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 $p(\boldsymbol{X}_t|\boldsymbol{Z}_t,\boldsymbol{A})$. In this case, we suggest that $q(\boldsymbol{x}_t|\boldsymbol{X}_{t-1},\boldsymbol{Z}_t,\boldsymbol{A}) = p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{A})$ is a suitable choice for the proposal function. This is similar to the common choice in 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. The weight π_t can then be shown as

$$\pi_t \propto \pi_{t-1} \cdot p(\boldsymbol{z}_t|\boldsymbol{x}_t). \tag{4}$$

We refer readers to [8] for the details of the derivation.

III. Mixture-Based AGPF

We have laid the theoretical foundation for AGPF and derived the probability propagation (3) that is related to the attractor-guided BN shown in Fig. 4. In addition, an importance sampling strategy is described for finding its approximate solution. We now consider how to set the state transition probability (or proposal function) $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{A})$, which is an important issue when employing the AGPF framework for tracking.

The transition probability can be set in various ways. We approximate it by setting $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{A})$ as a mixture distribution of $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_t|\boldsymbol{A}_i)$, $i = 1 \ldots n$

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{A}) = \alpha_0 \cdot p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) + \Sigma_{i=1,\ldots,n}\alpha_i \cdot p(\boldsymbol{x}_t|\boldsymbol{A}_i) \tag{5}$$

where $\Sigma_{i=0,\ldots,n}\alpha_i = 1$. In (5), the probability $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ is caused by the state transition, and the probability $p(\boldsymbol{x}_t|\boldsymbol{A})$ is caused by the attractors that regulate the state transition probability. We refer to this setting as a *mixture-based AGPF*. Using the mixture model in the transition probability does not have a strong physical background, but is based on a heuristic argument that is introduced to come to a feasible solution. Other nonmixture ways of specifying $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{A})$ are possible as long as such ways allow multimodal distributions. A characteristic of the mixture-based AGPF is that its particle set can be clearly separated into an online subset associated with the previous state $\boldsymbol{x}_{t-1}$ and an offline subset associated with the attractors $\boldsymbol{A}$, as analyzed below.

Since a sample set $\{\boldsymbol{s}_t^i, i = 1 \ldots N\}$ is generated from $\{\boldsymbol{s}_{t-1}^i\}$ via the proposal function $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{A})$ defined as a mixture of $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_t|\boldsymbol{A})$ in (5), 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 (5), 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 $\{\hat{\boldsymbol{s}}_t^k\} \cup \{\boldsymbol{a}^j\}$.

Note that, unlike $\{\hat{\boldsymbol{s}}_t^k, k = 1 \ldots M_1\}$, which have to be generated *online* from $\boldsymbol{x}_{t-1}$ during tracking, $\{\boldsymbol{a}^j; j = 1 \ldots M_2\}$, the particles from $\boldsymbol{A}$, can be constructed *offline* and stored before tracking since $\boldsymbol{A}$ are known in advance. More flexibly, the particles $\{\boldsymbol{a}^j\}$ can also be randomly chosen from a larger set of samples generated offline. The appearances of the pregenerated samples can be prerecorded for likelihood estimation; thus, our method combines particles related to both dynamic model-driven information and appearance information for tracking. Fig. 5 shows an illustration of the mixture-based AGPF. In our work, the re-sampling procedure [1] is also adopted as suggested in many particle filter-based methods.

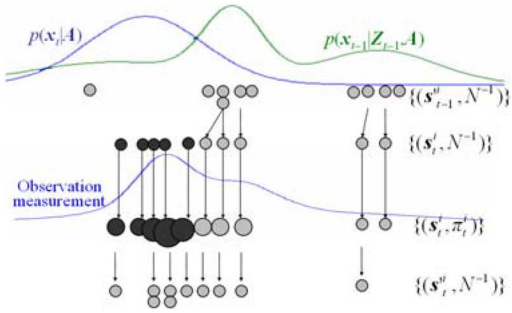


Fig. 5. Graphical representation of the mixture-based AGPF method at time step t. The black particles are generated from the attractors, i.e., $p(\boldsymbol{x}_t|\boldsymbol{A})$, and the gray particles are generated from the transition model according to the previous time steps, i.e., $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. Note that, for efficiency, the black particles can be generated "offline" via pseudo randomness.

Accordingly, in a mixture-based AGPF, some particles are generated from the attractors $\boldsymbol{A} = \{\boldsymbol{A}_1 \ldots \boldsymbol{A}_n\}$, and others are produced from the previous state, $\boldsymbol{x}_{t-1}$; whereas in the original particle filtering, all particles are taken from $\boldsymbol{x}_{t-1}$. Also, note that in [39], particles are generated on a linear approximation of the appearance manifold; thus, under tracking, the state is restricted to being synthesized from the precollected appearances. However, our method does not have this restriction and we allow the tracking to be performed in a free (state) space. For state space regions not covered by the precollected appearances in high-DOF tracking, our method can thus still work by tracking with the system dynamics. A more detailed relationship between mixture-based AGPF and particle filtering can be shown by substituting (5) into (3) as follows:

$$p(\boldsymbol{x}_t|\boldsymbol{Z}_t,\boldsymbol{A}) \propto p(\boldsymbol{z}_t|\boldsymbol{x}_t) \cdot \left[\Sigma_{i=1,\ldots,n}\alpha_i \cdot p(\boldsymbol{x}_t|\boldsymbol{A}_i) + \int \alpha_0 \cdot p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \cdot p(\boldsymbol{x}_{t-1}|\boldsymbol{Z}_{t-1},\boldsymbol{A}) d\boldsymbol{x}_{t-1}\right]. \tag{6}$$

The larger α_0 is, the more important the sequential motion transition information will be and vice versa. When α_0 is set to one, it degenerates to the original particle filtering in which only dynamically propagated information is used. In contrast, if α_0 is set to zero, only static information is used and our method degenerates to a pure appearance-based approach.

In our work, the probabilities $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ and $p(\boldsymbol{x}_t|\boldsymbol{A}_i)$ are modeled as Gaussian distributions

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

respectively, where Σ_1 and Σ_2 are diagonal covariance matrices. In practice, since the attractors far away from the current state usually have small probabilities that can be neglected, we only use the most significant K attractors, which have the largest probability values of $p(\boldsymbol{z}_t|\boldsymbol{x}_t = \boldsymbol{A}_i)$. As particles are generated from the local K attractors and the previous tracked state, the

Gaussian-mixture-based AGPF can be viewed as an instance of tracking by modeling the appearance locally and adaptively with a point (sample)-based representation. The appearance is locally approximated by both the pretracked state and some local attractors close to the pretracked state, and Bayesian probability propagation serves as a foundation of selecting appropriate sample points in the state space for the approximation.

IV. AGPF-Based Hand Tracking

A major difficulty with articulated hand tracking is that the motion DOF is too high, resulting in too many possible appearances. Although some approaches use a large number of precollected appearances [3], [29], collecting all the appearances for comparison would be infeasible. 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 recovered by gradually changing the articulated motion parameters. This is one reason that state-space methods (such as particle filtering) have been widely used in recent studies. However, since articulated configurations are only estimated from sequential transition information, state-space methods easily mis-track, especially when the motions between two consecutive images are large. In addition, once mis-tracking occurs, it cannot be corrected by later input images. Since the AGPF method integrates the motion transition model and appearance information, it avoids these difficulties.

For tracking applications, the likelihood $p\left(\boldsymbol{z}_t|\boldsymbol{x}_t=\boldsymbol{s}_t^i\right)$ has to be specified based on $\boldsymbol{x}_t$. As in many approaches [33], [39], an area of the skin color region and edge information are used to measure the likelihood for articulated hand tracking. Assume that these two visual clues are independent. Given $\boldsymbol{x}_t$, the probability $p(\boldsymbol{z}_t|\boldsymbol{x}_t)$ can be measured by the two factors, skin color area and edge information

$$p(\boldsymbol{z}_t|\boldsymbol{x}_t) \propto p_{\text{area}}(\boldsymbol{z}_t|\boldsymbol{x}_t) \cdot p_{\text{edge}}(\boldsymbol{z}_t|\boldsymbol{x}_t). \tag{7}$$

To evaluate $p_{\text{area}}(\boldsymbol{z}_t|\boldsymbol{x}_t)$, the 3-D model corresponding to $\boldsymbol{x}_t$ is projected onto the image plane to obtain a binary silhouette image $\boldsymbol{I}_t$, where $\boldsymbol{I}_t(i,j) = 1$ means the pixel (i,j) belongs to the projected silhouette; otherwise, $\boldsymbol{I}_t(i,j) = 0$. For $p_{\text{area}}(\boldsymbol{z}_t|\boldsymbol{x}_t)$, we compute the area difference, E, between $\boldsymbol{I}_t$ and $\boldsymbol{z}_t$ as

$$E = \sum_{0<i\le w, 0<j\le h} |\boldsymbol{I}_t(i,j) - \boldsymbol{C}(\boldsymbol{z}_t(i,j))| \tag{8}$$

where $\boldsymbol{C}$ is a pixel-wise skin-color classifier. The output value of $\boldsymbol{C}(\boldsymbol{z}_t(i,j))$ is set as one if $\boldsymbol{z}_t(i,j)$ is a foreground-color pixel; otherwise, it is set to zero if $\boldsymbol{z}_t(i,j)$ is a background pixel. In this work, the pixel-wise binary classifier, $\boldsymbol{C}$, is constructed by the method proposed by Jones and Rehg [21]. We then set

$$p_{\text{area}}(\boldsymbol{z}_t|\boldsymbol{x}_t) \propto \exp\left(\frac{-E^2}{2\sigma_1^2}\right) \tag{9}$$

where σ_1 is a standard deviation.

Note that when a particle is generated from $\boldsymbol{x}_{t-1}$ (i.e., $\boldsymbol{s}_t^i = \hat{\boldsymbol{s}}_t^k$ for some k), $\boldsymbol{I}_t$ is online synthesized from the corresponding 3-D hand model; otherwise, if it is generated from an attractor (i.e., $\boldsymbol{s}_t^i = \boldsymbol{a}^j$ for some j), $\boldsymbol{I}_t$ is the silhouette of a prestored appearance that is offline generated. Since projecting a 3-D hand model onto the image plane is time-consuming, applying several prestored silhouettes directly makes the implementation more efficient.

TABLE I
AGPF Algorithm for Articulated Hand Tracking

Offline step: For each attractor $\boldsymbol{A}_i$ ($i = 1,\ldots, n$), a set of particles is pre-generated and stored.

Tracking algorithm: Given a set of weighted samples $\{(\boldsymbol{s}_{t-1}^i, \pi_{t-1}^i),\ i = 1,\ldots, N\}$ at time step t–1, the following steps are performed to construct a new set of samples at time step t.

1. Re-sample a particle set $\{(\boldsymbol{s'}_{t-1}^i, N^{-1}),\ i = 1\ldots N\}$ from the particles $\{(\boldsymbol{s}_{t-1}^i, \pi_{t-1}^i), i = 1,\ldots, N\}$.
2. Sample a particle set $\{\boldsymbol{s}_t^i, i = 1\ldots N\}=\{\hat{\boldsymbol{s}}_t^k\}\cup\{\boldsymbol{a}^j\}$ as follows:
   - 2.1. Choose the most significant K attractors ($K<n$) having the largest probability values of $\{p(\boldsymbol{z}_t|\,\boldsymbol{x}_t = \boldsymbol{A}_i), i = 1,\ldots, n\}$.
   - 2.2. Randomly select M_1 particles $\{\boldsymbol{s'}_{t-1}^k,\ k = 1,\ldots, M_1\}$ from $\{\boldsymbol{s'}_{t-1}^i\}$ where $M_1 = \alpha_0 \cdot N$ ($0 \le \alpha_0 \le 1$). Then, online generate $\{\hat{\boldsymbol{s}}_t^k,\ k = 1,\ldots, M_1\}$ based on the transition model $p(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1} = \boldsymbol{s'}_{t-1}^k)$.
   - 2.3. Randomly select M_2 particles, $\{\boldsymbol{a}^j, j = 1,\ldots, M_2\}$, from the pre-generated particles of the K attractors, where $M_2 = (1-\alpha_0)\cdot N$.
3. Measure the weight π_t^i of sample $\boldsymbol{s}_t^i$ based on the likelihood $p(\boldsymbol{z}_t|\,\boldsymbol{x}_t=\boldsymbol{s}_t^i)$ in (7).
4. Estimate the state $\boldsymbol{x}_t^*$ by using (11) for the display.

To estimate $p_{\text{edge}}(\boldsymbol{z}_t|\boldsymbol{x}_t)$, we use the directed Chamfer distance (DCD) [4], 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. Given two sets of edge points, $\boldsymbol{U}$ and $\boldsymbol{V}$, obtained from the contours of $\boldsymbol{I}_t$ and $\boldsymbol{C}(\boldsymbol{z}_t)$, respectively, the likelihood of an edge is defined as

$$p_{\text{edge}}(\boldsymbol{z}_t|\boldsymbol{x}_t) \propto \exp\left(\frac{-D^2(U,V)}{2\sigma_2^2}\right) \tag{10}$$

where $D(\boldsymbol{U},\boldsymbol{V})$ is the Chamfer distance between $\boldsymbol{U}$ and $\boldsymbol{V}$, and σ_2 is another standard deviation. In our work, the distance transform [16] is used for efficient computation of the DCD.

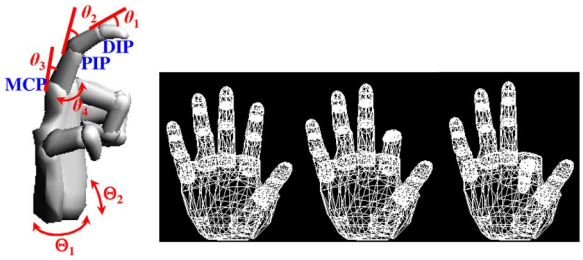
The weight π_t^i of each sample in $\left\{\boldsymbol{s}_t^i\right\}$ is calculated from the observation distribution $p\left(\boldsymbol{z}_t|\boldsymbol{x}_t=\boldsymbol{s}_t^i\right)$ based on (7). Thus, the desired posterior distribution $p(\boldsymbol{x}_t|\boldsymbol{Z}_t, \boldsymbol{A})$ can be represented by the set of weighted samples $\left\{\left(\boldsymbol{s}_t^i, \pi_t^i\right)\right\}$. The state $\boldsymbol{x}_t^*$ at time step t for the display is estimated by the maximum mode, as suggested in [13]

$$\boldsymbol{x}_t^* = \boldsymbol{s}_t^*, \text{ where } \pi_t^* = \max\left(\pi_t^i\right). \tag{11}$$

We summarize the algorithm in Table I.

V. Experimental Results of Hand Tacking

The generic 3-D hand model used in the experiments is shown in Fig. 6(a). It has been used in a number of works [33], [39]. There are two global rotations; Θ_1 and Θ_2. Each



(a) (b)

Fig. 6. (a) Hand structure and articulations. (b) Three selected states of the index finger.

finger has three joints, the MCP, PIP, and DIP, which have two degrees, one degree, and one degree of freedom (DOF), respectively. In this work, we assume that there is no *a priori* knowledge about the hand gestures in the test sequence; that is, all the hand articulations have an equal possibility of appearing. Therefore, to deal with the input sequence in a general hand-tracking problem, we select attractors that are uniformly distributed in the state space. Since one of the rotations of the MCP joint, θ_4, is usually small, we neglect it when selecting the attractors. Then, only three angles $(\theta_1, \theta_2, \theta_3)$ are considered. For each finger, we coarsely sample the three states $\{(0,0,0),(0,90^\circ,0),(0,0,90^\circ)\}$, as shown in Fig. 6(b). The sampled states of different fingers are combined to form the attractors. Hence, there are 3^m $(m \leq 5)$ attractors if m fingers are involved in the tracking. In addition, nine states associated with Θ_1 and Θ_2 are precollected for global hand motion by rotating the palm at different angles, so there are 9×3^m attractors in the experiments when global rotations are considered.

To choose the most significant K attractors efficiently, a complete graph is constructed in the state space beforehand, on which the attractors are the nodes and the edge cost is the distance between nodes in the state space. Instead of computing $p(\boldsymbol{z}_t|\boldsymbol{x}_t = \boldsymbol{A}_i)$ for all i, only a limited number of attractors, which have the minimum edge costs for the selected attractors at time $t-1$, are evaluated for selecting the most significant K attractors.

We perform a series of experiments on both simple and cluttered backgrounds. Image sequences with a large range of motions are captured, and the image resolution is 320×240. Two hundred particles (i.e., $N = 200$) are employed for each experiment, and α_i is set to $1/(K+1)$, for $i = 0, \ldots, K$ (K is set to 2 in our experiments). The elements of diagonal matrices Σ_1 and Σ_2 are set to 70 (degrees^2). We compare the tracking results of the AGPF method with those obtained by using standard particle filtering only, and by using appearance information only. Images are registered along one of the sides of the foreground region's bounding box which has the maximum likelihood.

In the first experiment (Fig. 7), a 14-DOF model obtained by bending three fingers and rotating the palm with 2 DOFs is used. There are $9 \times 3^3 = 243$ attractors. Fig. 7(a) shows part of the input images. The tracking results of standard particle filtering, the pure appearance-based approach (by choosing an attractor $\boldsymbol{A}_i$ with the maximum probability $p(\boldsymbol{z}_t|\boldsymbol{x}_t = \boldsymbol{A}_i)$), and the AGPF method are shown in Fig. 7(b)–(d), respectively. One can see that simple particle filtering fails to track in several images. Moreover, the tracking performance is poor when pure appearance information is used. In contrast, the proposed AGPF method recovers these motions, as shown in Fig. 7(d).

To quantify the proposed method, numerical evaluations are also performed. Since the ground truth of hand articulation in real video sequences is difficult to obtain, we measure the errors in the image space to evaluate the performance approximately. In our test, the hand region of each input image is segmented manually in advance. The area difference between the presegmented hand region and the projected silhouette of an estimated state is then computed, and the ratio of the area difference to the hand area serves as the error measurement for the evaluation. Note that the articulated 3-D hand model used in our work is a generic one, which does not match the hands in the images of our experiments exactly. Nevertheless, the ratio defined still serves as a relatively accurate measurement to quantify the results. The evaluation results of this experiment are shown in the second column of Table II. The average area-difference ratios of standard particle filtering, the pure appearance-based approach, and the AGPF method are 0.229, 0.277, and 0.201, respectively. In this evaluation, it shows that AGPF outperforms the other two methods.

Next, we use the above experiment to demonstrate how the tracking trajectory is influenced by the attractors. Fig. 8 shows the tracking trajectories of simple particle filtering and AGPF, where principal component analysis (PCA) is used to reduce the high-dimensional representation of samples to a 2-D space for visualization. The blue points and line represent, respectively, some of the samples and the trajectory generated by simple particle filtering; the red points and line represent those generated by the AGPF method. The point sign and cross sign represent the samples generated by the motion transition model. The star sign, $*$, represents the samples generated by attractors. In addition, the attractor $\boldsymbol{A}_i$ with the maximum probability, $p(\boldsymbol{z}_t|\boldsymbol{x}_t = \boldsymbol{A}_i)$, is shown as a green circle. From this figure, we observe that the attractors can guide the tracker toward a distinguishing motion trajectory in the state space.

In the second experiment (Fig. 9), a 16-DOF model by bending four fingers and $3^4 = 81$ attractors are used. As with the first experiment, the tracking results using simple particle filtering or pure appearance information are not very accurate in this experiment either. However, our method provides better results, as shown in Fig. 9(d) and Table II.

Some other experiments were performed with different DOFs and different numbers of attractors. Fig. 10 shows the results of tracking four fingers, where the DOF is 16 with 81 attractors, while Fig. 11 shows the tracking results of global hand motion with five-finger articulation, where the DOF is 22 with 2,187 attractors. The fourth and the fifth columns of Table II show their numerical results, respectively. The experiment results demonstrate that, with a limited number of precollected appearances, our approach significantly refines the performance of particle filtering in which only a motion transition model is used, and also performs better than cases when only appearances are used.

VI. AGPF-Based Lip-Contour Tracking

Although we used articulated hand tracking as an example in the above discussion, AGPF serves as a general framework for

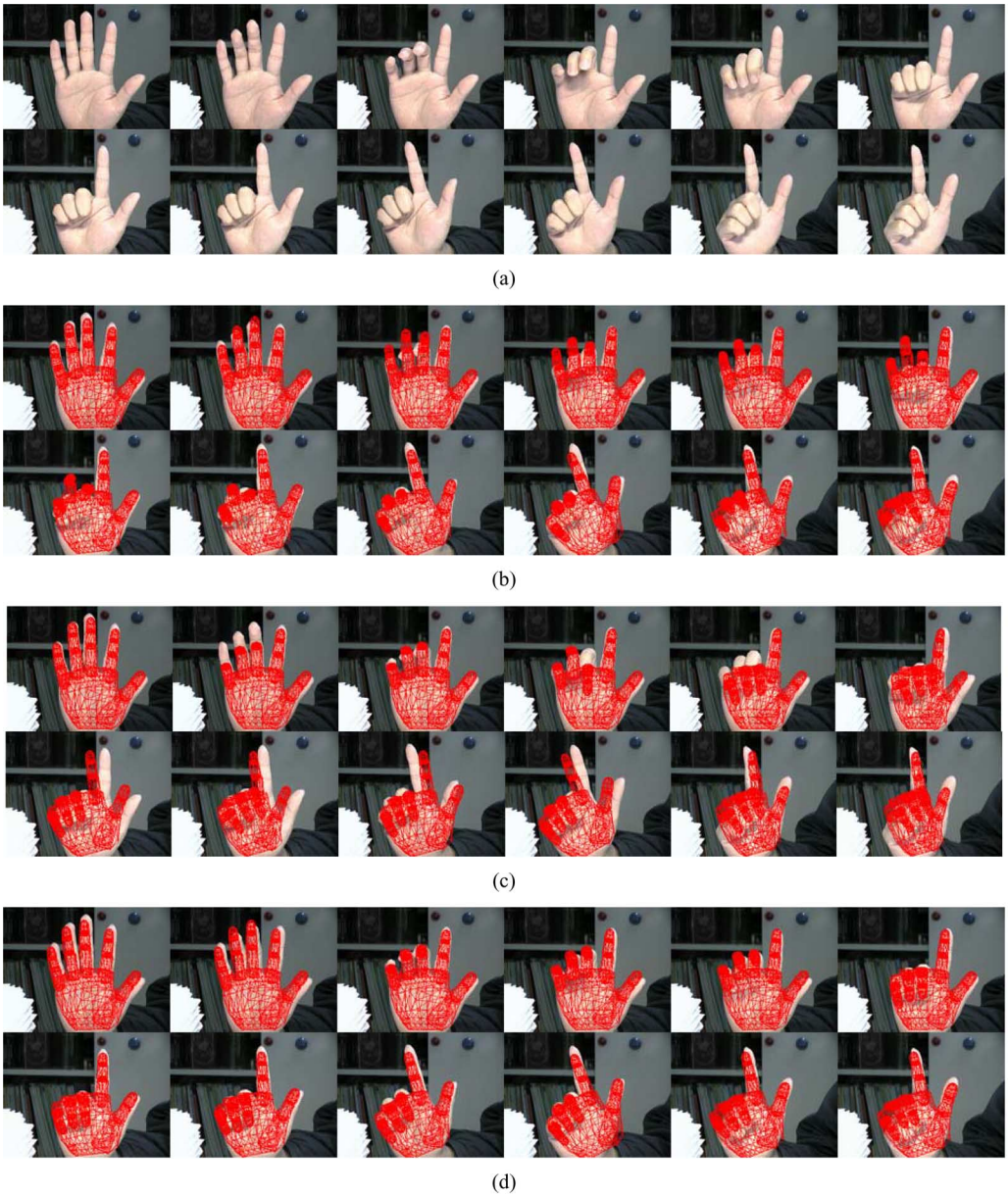


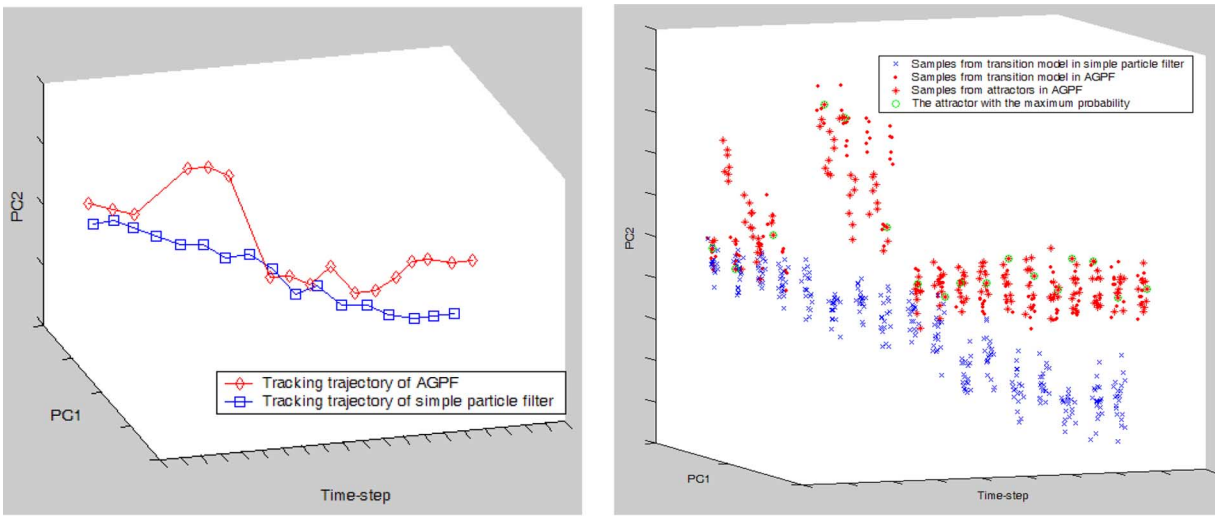
Fig. 7. Global hand motion with three-finger articulation in 14 DOFs: (a) input sequence, (b) tracking results using simple particle filtering, (c) tracking results using appearance information only, and (d) tracking results using the AGPF method.

TABLE II
AVERAGE AREA-DIFFERENCE RATIOS OF THE ARTICULATED HAND-TRACKING EXPERIMENTS

	Exp 1 (Fig. 7)	Exp 2 (Fig. 9)	Exp 3 (Fig. 10)	Exp 4 (Fig. 11)
Simple Particle Filtering	0.229	0.194	0.221	0.356
Pure Appearance-based	0.277	0.203	0.226	0.438
AGPF	**0.201**	**0.188**	**0.211**	**0.308**

appearance-guided state estimation of a dynamic system. It is also applicable to other high DOF model-based tracking or contour-based tracking applications. In this section, we apply AGPF to lip-contour tracking. Though the problem complexity is relatively lower than that of articulated hand tracking, our objective is to demonstrate that AGPF is effective for tracking applications of both simple and complex cases, as long as the nature of the problem allows a state-observation mapping relationship (i.e., attractors with known observations) to be established in advance.

Due to the smoothness and elasticity constraints, active contours [22] are popular for lip-contour tracking. However, the unclear boundary between lip and facial skin makes the snakes unreliable. It is also difficult to tune the parameters of the snakes. To track the lip-contour robustly, methods based on *a priori* shape knowledge have been suggested recently.



(a) (b)

Fig. 8. Tracking trajectories reduced to a 2-D PCA space (PC1×PC2) in association with simple particle filtering and the AGPF of the first experiment (Fig. 7), where PC1 and PC2 are the first and second principle components, respectively. (a) The trajectories of these two methods. (b) The samples generated from these two methods.

The active shape model (ASM) [12] and the active appearance model (AAM) [11] are two successful methods in this field. Although the ASM/AAM can provide relatively convincing results, they share some limitations with common appearance-based approaches: A large training set is necessary to obtain the eigen-bases for covering the lip shape variability generally and the control points of the training set have to be carefully labeled in advance. To avoid these limitations, we adopt the AGPF method for lip-contour tracking. Limited shape priors are precollected and treated as attractors in AGPF-based lip-contour tracking.

To reduce the labeling burden, we adopt the radial vector model [9] as shown in Fig. 12 to represent the contour of a lip in AGPF-based lip-contour tracking. The radial vectors are uniformly spread over 360°, and each vector originates from the centroid of the contour and links to a contour point. The shape of the contour is deformed by varying the distances $\{l_1, l_2, \ldots, l_n\}$ of the radial vectors, and the centroid of the contour moves during the deformation. The angular interval θ controls the smoothness of the contour and the number of control points in the lip contour is equal to $n = 360/\theta$. There are some advantages in using this representation. We can avoid the effort of labeling control points and easily control the dimension of state variables to compromise between the tracking time and the visual results.

With the radial vector model, we define the state vector as $\boldsymbol{x}_t = (L_t, c_t)$, where $L_t = \{l_1, l_2, \ldots, l_n\}$ represents the n radial vector at time t and c_t is the centroid of the lip. A set of precollected shape priors represented by the radial vector model, is treated as attractors. Besides generating particles by using (5), uniform sampling scheme can also be adopted [19]. To distinguish the regions of facial skin and the lip, a discriminative feature representation is required. In our method, we adopt the feature selection algorithm proposed by Collins *et al.* [10] to obtain a linear color projection function $\omega = (\omega_1, \omega_2, \omega_3)$ such that the feature value has high discriminability between the color of the facial skin and that of the lip. For a pixel I_u, its feature value F is computed by its RGB values and the color projection function ω, i.e., $F \equiv \langle I_u, \omega \rangle \equiv \omega_1$ R+ω_2G+ωB. Given a color projection function ω, let $H_{\text{lip}}(i)$ be a histogram of the feature value for pixels inside the lip contour, and $H_{bg}(i)$ be a histogram for pixels from outside that contour, where index i ranges from 1 to b, the number of histogram bins. Then, discrete probability densities, $P_{\text{lip}}(\cdot)$ for the lip and $P_{bg}(\cdot)$ for the background, are obtained by normalizing each histogram. The variance ratio (VR) defined in [10] is used to quantify the difference between $P_{\text{lip}}(\cdot)$ and $P_{bg}(\cdot)$ under feature F

$$VR(G; P_{\text{lip}}, P_{bg}) \equiv \frac{\text{var}\left(G; \frac{P_{\text{lip}}+P_{bg}}{2}\right)}{\text{var}(G; P_{\text{lip}}) + \text{var}(G; P_{bg})} \tag{12}$$

where

$$\begin{cases} G(i) = \log \frac{\max\{P_{\text{lip}}, \varepsilon\}}{\max\{P_{bg}, \varepsilon\}} \\ \text{var}(G; h) = \sum_i h(i) G^2(i) - \left[\sum_i h(i) G(i)\right]^2 \end{cases} \tag{13}$$

and ε is a small positive value that ensures the denominator is nonzero.

The likelihood $p(\boldsymbol{z}_t|\boldsymbol{x}_t)$ is defined as being proportional to the VR between the interior and exterior histograms

$$p(\boldsymbol{z}_t|\boldsymbol{x}_t) \propto 1 - \exp\left(\frac{-VR(G; P_{\text{lip}}, P_{bg})^2}{2\sigma_3^2}\right) \tag{14}$$

where σ_3 is a standard deviation. Fig. 13 shows a lip image and its likelihood image via a color projection function.

To demonstrate AGPF's performance in lip-contour tracking, two sequences with 640×480 resolutions are used in our experiments. Two hundred particles are employed in these experiments and the weight α_i is also set to $1/(K+1)$ where $K = 3$. The first n elements of diagonal covariance matrices, Σ_1 and

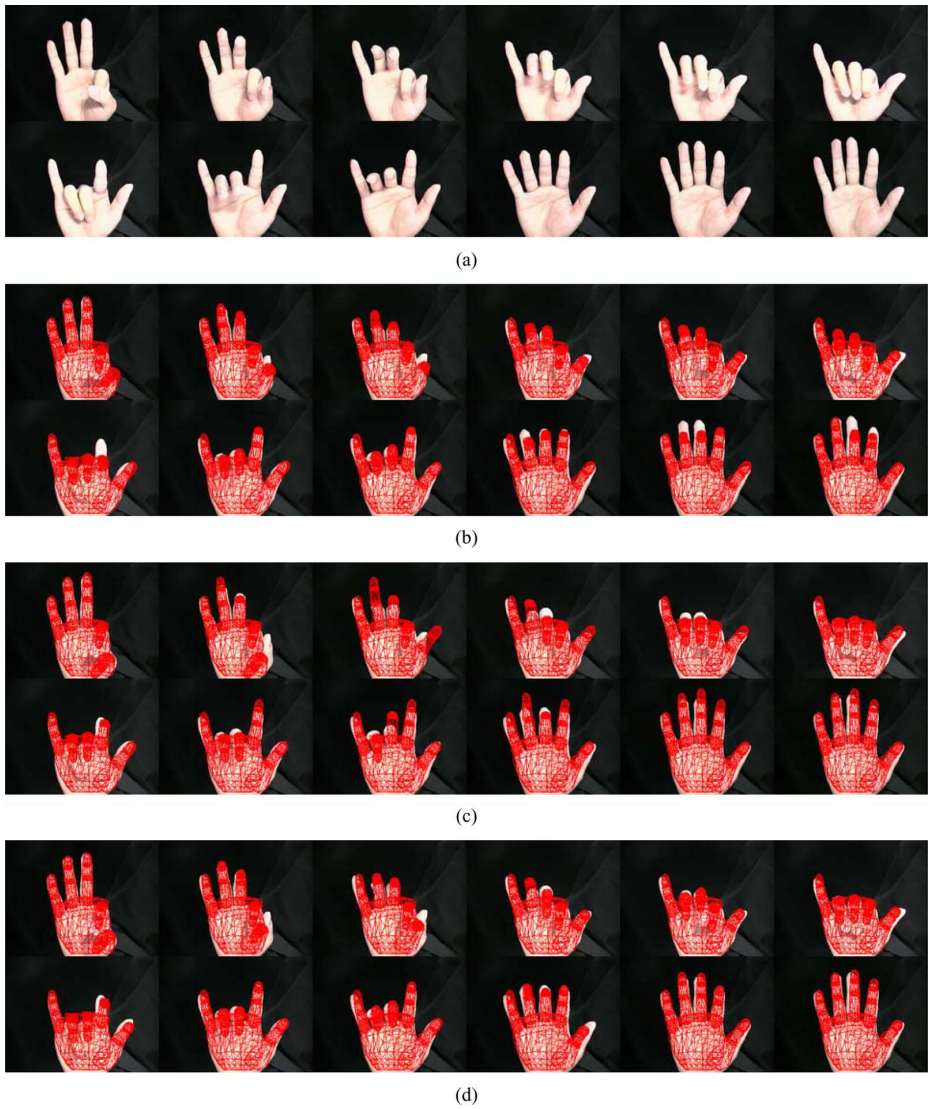


Fig. 9. Four-finger articulation with 16 DOFs: (a) input sequence, (b) tracking results using simple particle filtering, (c) tracking results using appearance information only, (d) tracking results using the AGPF method.

Σ_2, represent the variances of the distances between the centroid and contour. In the experiments, we set these elements as 9 (pixel^2). The last two elements of the diagonal covariance matrices represent the 2-D motion variances of the centroid, and we set them as 4 (pixel^2). The angular interval θ of the radial vector model is 20°; thus, we have 18 control points on the lip contour and the DOF of lip-contour tracking is 20, including the 2-D position of the centroid of the contour. The number of bins for the lip and background histograms is set as 32. An algorithm similar to that shown in Table I is also performed by using the likelihood measurement in (14).

As with the experiments on articulated hand tracking, we also compare the tracking results of AGPF with those by using particle filtering and using appearance/attractor information only. Figs. 15 and 16 show the experimental results of lip-contour tracking for different people using ten and eight attractors, respectively, and some of the attractors are shown in Fig. 14. Since the shape variation of lip-contours is not as large as that of articulated hand tracking, the attractors are not distributed over the state space. Instead, they are selected empirically based on contours, which are easily mis-tracked in particle filtering. The color projection function ω used in these experiments is $\omega = (1, -2, 1)$. By computing the difference of the area inside the tracked lip-contour and the ground truth labeled manually, errors in the area-difference ratio are also measured. In the experiment associated with Fig. 15, the area-difference ratios of the method only using attractor information and our AGPF method are 0.175 and 0.141, respectively, while the tracker using simple particle filtering begins to drift in the 75th frame (out of a total of 300 frames), and its error is far larger. In this application, AGPF again outperforms the other methods. The experimental results demonstrate that the AGPF framework can accurately estimate the contour between the lip and the facial skin.

VII. Discussion

The AGPF framework introduced in this paper incorporates appearance information into state-space tracking to form a combined approach. In AGPF, we assume that a mapping

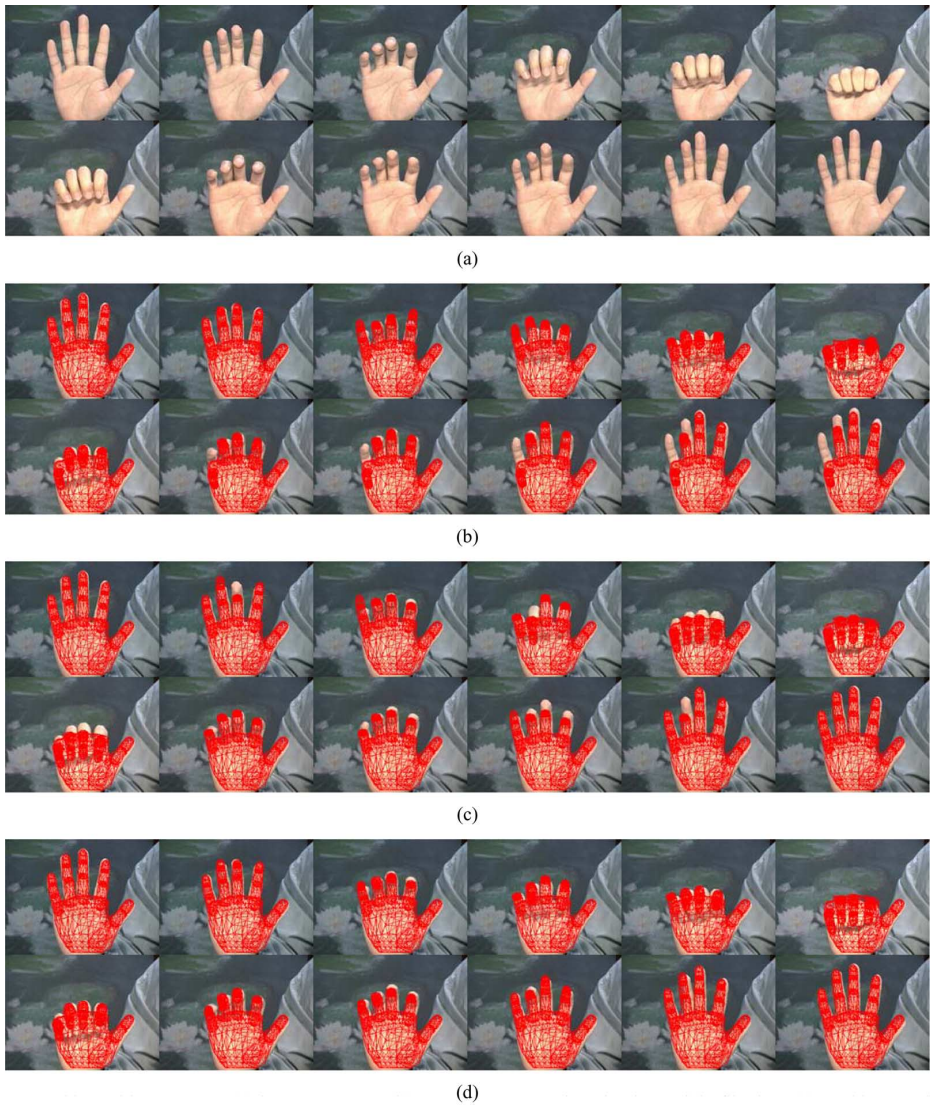


Fig. 10. Four-finger tracking with 16 DOFs: (a) input sequence, (b) tracking results using simple particle filtering, (c) tracking results using appearance information only, and (d) tracking results using the AGPF method.

from the state space to the observation space is available, or equivalently, the observation is a function of the state to be estimated. It is clear that, without this assumption, the concept of attractors can not be clearly defined, and AGPF can not be well formulated.

AGPF is not suitable for a few tracking problems about which the above assumption cannot be made. For example, when the tracking problem simply involves locating the position of a target in an image, the state (i.e., image position) and the observation (i.e., the target appearance) usually can not be formulated by a unique preknown mapping. In this case, to incorporate appearance information into particle filtering, some approaches treat the appearance as another set of variables (or states) to be estimated with Rao-Blackwellized particle filtering [23]. However, this is not an effective way to employ appearance information when the state-observation mapping can be clearly identified. On the other hand, AGPF is more suitable for such situations.

For tracking problems that AGPF can be applied to, we have shown that the method can improve the performance of particle filtering or pure appearance-based matching. This is because static attractors pregathered in the state space can serve as effective guides to regulate the system dynamics and prevent drifting. Two examples, 3-D articulated hand tracking and lip-contour tracking, have been studied to verify AGPF's effectiveness.

Another issue worth noting is that the proposed method can be interpreted in several ways. According to (3), it seems that we only need to set the proposal function $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ in standard particle filtering based on the precollected attractors $\boldsymbol{A}$, then standard particle filtering can be re-formulated in the same way as AGPF. This is correct from the implementation point of view, but we wish to emphasize that this approach can be studied more formally by the attractor-guided BN in Fig. 4. While the solution of the BN can be derived such that it has an optimal belief-propagation form similar to that of particle filtering, we have a clearer insight into how the attractors $\boldsymbol{A}$ can be chosen appropriately to regulate the system dynamics.

Since the mapping between the state space and the observation space is known in our framework, fixed appearance in-

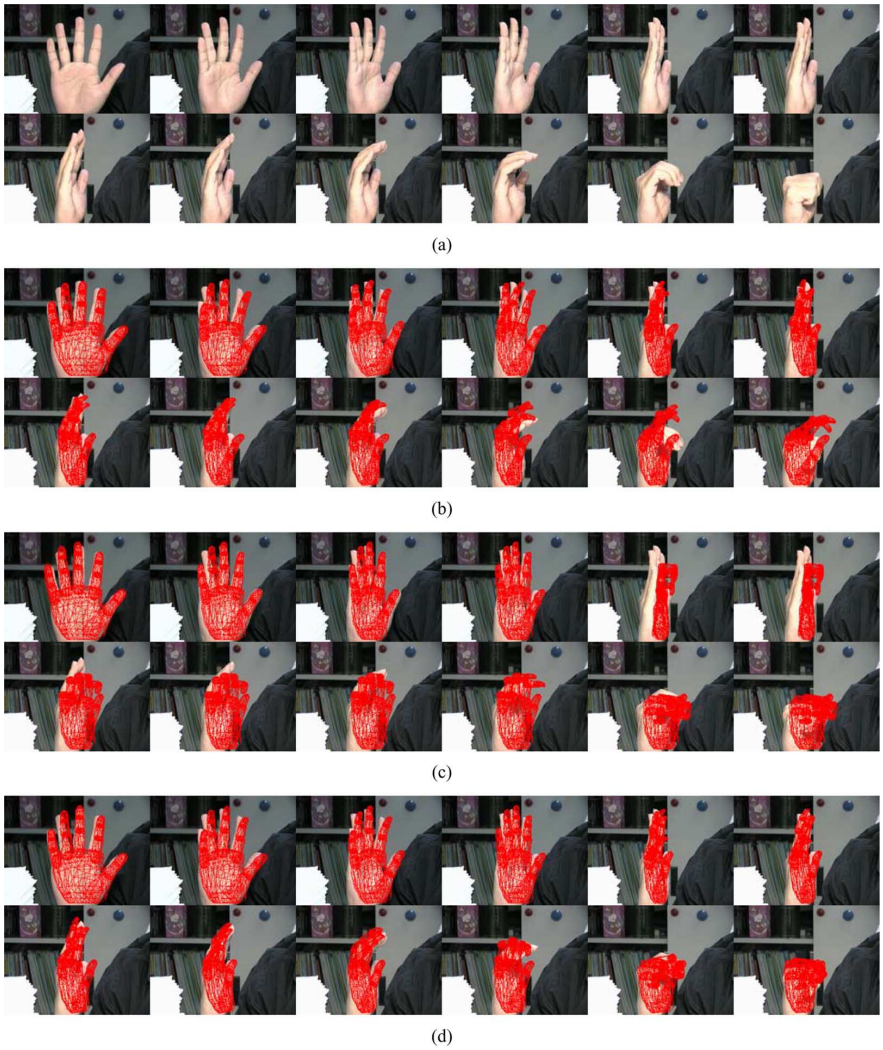


Fig. 11. Global hand motion with five-finger articulation with 22 DOFs: (a) input sequence, (b) tracking results using simple particle filtering, (c) tracking results using appearance information only, and (d) tracking results using the AGPF method.

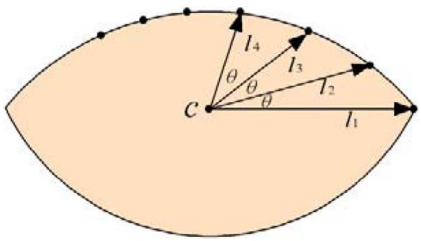


Fig. 12. Radial vector model, which decomposes the lip contour into $360/\theta$ control points; c is the centroid of the lip contour and l_i is the distance of the ith radial vector from the centroid to a contour point.

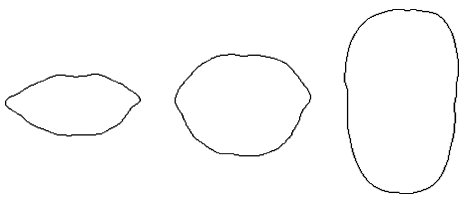


Fig. 14. Some attractors used in lip-contour tracking. From left to right, the attractors represent the shapes of closed, half-open and fully-open lips, respectively.

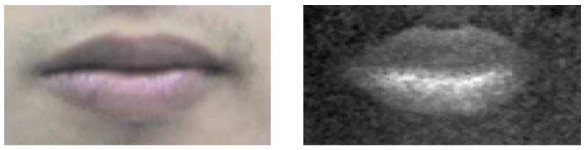


Fig. 13. Likelihood image. (a) Lip image. (b) Likelihood image of (a), obtained by using the method in [10].

formation is adopted for the presented applications. Alternatively, the appearance information (or attractors) can be trained over time. This method is used primarily for applications whose state-observation mapping cannot be clearly identified in advance. For example, in [40], the target appearance tracked in the previous time step is used to update the online appearance model sequentially. Okuma *et al.* [27] used Adaboost to generate attractors online and incrementally updated the attractors over time. In addition, how to design a good proposal function is the main issue that arises when using particle filtering for various applications [13], [15], [18], [20], [27], [30], [37], [39]. Our dynamic BN explanation provides a new and general way of looking at this type of approach.

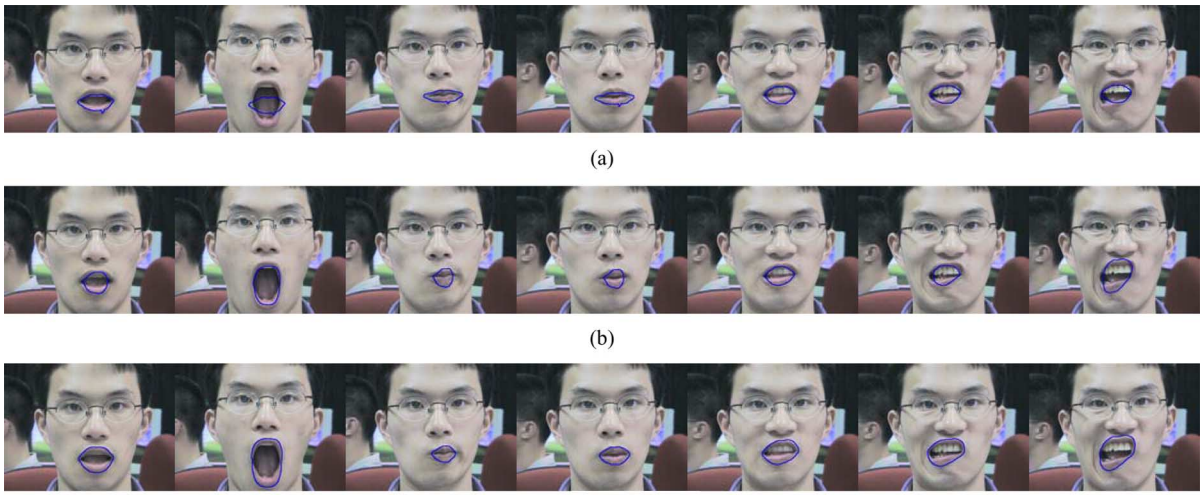


Fig. 15. Lip-contour tracking: (a) tracking results using simple particle filtering, (b) tracking results using attractor information only, and (c) tracking results using the AGPF method.

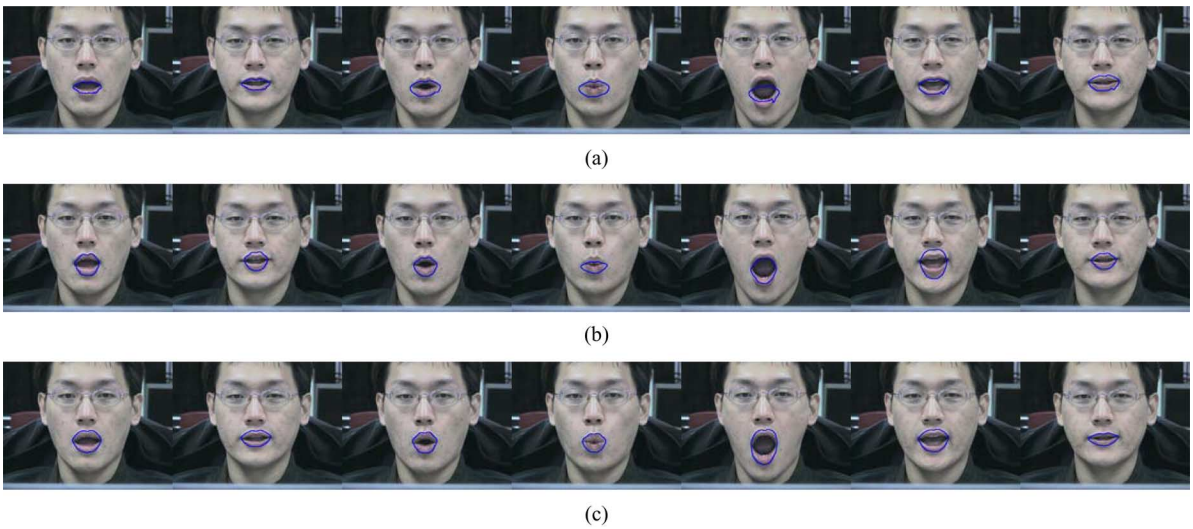


Fig. 16. Lip-contour tracking for different people: (a) tracking results using simple particle filtering, (b) tracking results using attractor information only, and (c) tracking results using the AGPF method.

VIII. Conclusion

In this paper, we have presented a model-based tracking framework that incorporates static appearance information into a dynamic system. We derive the Bayesian probability propagation of the MAP solution when known attractors in the state space are involved in the system, and introduce a particle filtering approach to find an approximation of the MAP solution. By representing the transition probability as a mixture distribution, we introduce the mixture-based AGPF, a systematic tracking approach in which the appearance manifold is locally re-modeled via point-based approximation during tracking. Our approach avoids the drifting effect of particle filtering by using a limited number of precollected attractors to guide the tracking in a high-dimensional state space. We also allow the tracking to be performed via the system dynamics in free space when complete appearance information is difficult to collect for a high-DOF tracking problem. The proposed method yields promising results in applications of articulated hand tracking and lip-contour tracking.

Currently, the attractors of the AGPF Bayesian network (Fig. 4) are selected empirically or by uniformly distributing them in the state space. However, for applications in which specific motion sequences are targeted, our approach can be modified to that of selecting a set of particular attractors in advance through prelearning, where the targeted sequences can serve as the training sequences. The attractors can then be set as the motion states easily drift in the prelearning stage. In the future, we will investigate how to select the attractors by a prelearning process in order to boost the tracking accuracy and efficiency of applications in which particular motion sequences are targeted.

Acknowledgment

The authors would like to thank the anonymous reviewers and the associate editor for their valuable comments.

REFERENCES

[1] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nolinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[2] V. Athitsos and S. Sclaroff, "An appearance-based framework for 3D hand shape classification and camera viewpoint estimation," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2002, pp. 40–45.

[3] V. Athitsos and S. Sclaroff, "Estimating 3D hand pose from a cluttered image," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, vol. 2, pp. 432–439.

[4] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," in *Proc. Int. Joint Conf. Artificial Intelligence*, 1977, pp. 659–663.

[5] A. G. Bors and I. Pitas, "Prediction and tracking of moving objects in image sequences," *IEEE Trans. Image Process.*, vol. 9, no. 8, pp. 1441–1445, Aug. 2000.

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

[7] M. Carcassoni and E. R. Hancock, "Correspondence matching with modal clusters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1609–1615, Dec. 2003.

[8] W. Y. Chang, C. S. Chen, and Y. P. Hung, "Appearance-guided particle filtering for articulated hand tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 235–242.

[9] G. I. Chiou and J. N. Hwang, "Lipreading from color video," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1192–1195, Aug. 1997.

[10] R. T. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1631–1643, Oct. 2005.

[11] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[12] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models: Their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.

[13] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture an annealed particle filtering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000, vol. 2, pp. 2126–2133.

[14] 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.

[15] P. Elinas, R. Sim, and J. J. Little, "SigmaSLAM: Stereo vision SLAM using the Rao-Blackwellised particle filter and a novel mixture proposal distribution," presented at the IEEE Int. Conf. Robotics and Automation, 2006.

[16] R. Haralick and L. Shapiro, *Computer and Robot Vision*. Reading, MA: Addison-Wesley, 1992, vol. 1.

[17] M. Isard and A. Blake, "CONDENSATION-Conditional density propagation for visual tracking," *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.

[18] M. Isard and A. Blake, "ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework," in *Proc. Eur. Conf. Computer Vision*, 1998, pp. 893–908.

[19] Y. D. Jian, W. Y. Chang, and C. S. Chen, "Attractor-guided particle filtering for lip contour tracking," in *Proc. Asian Conf. Computer Vision*, 2006, vol. 1, pp. 653–663.

[20] Y. Jin and F. Mokhtarianm, "Towards robust head tracking by particles," in *Proc. IEEE Int. Conf. Image Processing*, 2005, vol. 3, pp. 864–867.

[21] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Int. J. Comput. Vis.*, vol. 46, no. 1, pp. 81–96, 2002.

[22] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, 1988.

[23] Z. Khan and T. B. Dellaert, "A Rao-Blackwellized particle filter for eigentracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 980–986.

[24] B. Li and R. Chellappa, "A generic approach to simultaneous tracking and verification in video," *IEEE Trans. Image Process.*, vol. 11, no. 5, pp. 530–544, May 2002.

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

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

[27] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *Proc. European Conf. Computer Vision*, 2004, pp. 28–39.

[28] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 677–695, Jul. 1997.

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

[30] Y. Rui and Y. Chen, "Better proposal distributions: Object tracking using unscented particle filter," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, vol. 2, pp. 786–793.

[31] S. Russell and P. Norvig, *Artificial Intelligence-A Modern Approach*. Englwood Cliffs, NJ: Prentice-Hall, 1995.

[32] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter-sensitive hashing," in *Proc. IEEE Int. Conf. Computer Vision*, 2003, vol. 2, pp. 750–757.

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

[34] E. Sudderth, M. Mandel, W. Freeman, and A. Willsky, "Distributed occlusion reasoning for tracking with nonparametric belief propagation," presented at the Advances in Neural Information Processing Systems, 2004.

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

[36] J. Triesch and C. v. d. Malsurg, "A system for person-independent hand posture recognition against complex background," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1449–1453, Dec. 2001.

[37] Q. Wen, J. Gao, A. Kosaka, H. Iwaki, K. Luby-Phelps, and D. Mundy, "A particle filter framework using optimal importance function for protein molecules tracking," in *Proc. IEEE Int. Conf. Image Processing*, 2005, vol. 1, pp. 1161–1164.

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

[39] Y. Wu, J. Y. Lin, and T. S. Huang, "Analyzing and capturing articulated hand motion in image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1910–1922, Dec. 2005.

[40] S. K. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Trans. Image Process.*, vol. 13, no. 12, pp. 1491–1506, Dec. 2004.

Wen-Yan Chang received the B.S. degree in computer and information science from Tunghai University, Taichung, Taiwan, R.O.C., in 1998, and the M.S. degree in computer science and information engineering from the National Cheng Kung University, Tainan, Taiwan, in 2000. He is currently pursuing the Ph.D. degree in computer science and information engineering at the National Taiwan University, Taipei.

He is currently a Research Assistant in the Institute of Information Science, Academia Sinica, Taipei. His research interests include image processing, computer graphics, pattern recognition, and computer vision.

Chu-Song Chen received the B.S. degree in control engineering from the National Chiao-Tung University, Hsing-Chu, Taiwan, R.O.C., in 1989, and the M.S. and Ph.D. degrees from the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, in 1991 and 1996, respectively.

He is now an Associate Research Fellow of the Institute of Information Science, Academia Sinica, Taiwan, and also an Adjunct Associate Professor of the Graduate Institute of Networking and Multimedia, National Taiwan University. He has published more than 70 technical papers. His research interest includes pattern recognition, computer vision, signal/image processing, and multimedia.

Dr. Chen has served as the Secretary-General of the Image Processing and Pattern Recognition (IPPR) Society, Taiwan, since 2007, which is one of the societies of the International Association of Pattern Recognition (IAPR). He received the outstanding paper awards of IPPR in 1997, 2001, and 2005.

Yong-Dian Jian received the B.S.E. and M.S. degrees in computer science from the National Taiwan University, Taipei, Taiwan, R.O.C., in 2002 and 2004, respectively.

He is currently a research assistant at Institute of Information Science, Academia Sinica, Taiwan. His research interests include computer vision and machine learning.