Capturing Natural Hand Articulation

Ying Wu, John Y. Lin and Thomas S. Huang
Beckman Institute
University of Illinois at Urbana-Champaign
405 N. Mathews, Urbana, IL 61801
{yingwu,jy-lin,huang}@ifp.uiuc.edu

Abstract

Vision-based motion capturing of hand articulation is a challenging task, since the hand presents a motion of high degrees of freedom. Model-based approaches could be taken to approach this problem by searching in a high dimensional hand state space, and matching projections of a hand model and image observations. However, it is highly inefficient due to the curse of dimensionality. Fortunately, natural hand articulation is highly constrained, which largely reduces the dimensionality of hand state space. This paper presents a model-based method to capture hand articulation by learning hand natural constraints. Our study shows that natural hand articulation lies in a lower dimensional configurations space characterized by a union of linear manifolds spanned by a set of basis configurations. By integrating hand motion constraints, an efficient articulated motion-capturing algorithm is proposed based on sequential Monte Carlo techniques. Our experiments show that this algorithm is robust and accurate for tracking natural hand movements. This algorithm is easy to extend to other articulated motion capturing tasks.

1 Introduction

The use of hand gestures has become an important part of human computer interaction in recent years [11]. Gesture commands could be captured and recognized by computers, and computers could synthesize hand sign language as an output. Glove-based devices have been employed to capture human hand motion by directly measuring the joint angles and spatial positions of hands with sensors attached. Generally, such devices are expensive and cumbersome. On the other hand, vision-based technique has become a promising alternative to capture human hand motion, due to the cost-efficient and non-invasive visual sensory inputs. It serves as the motivating forces for research in vision-based capturing of hand articulation.

Capturing hand articulation is a challenging task, since the hand presents a motion of high degrees of freedom. If hand articulation is represented by its joint angles, the dimensionality for estimation and tracking of hand states would make this task prohibitive. Another difficulty comes from self-occlusions of different fingers, which brings uncertainty for the occluded parts.

Two general approaches have been explored to capture hand articulation. One of them is the *model-based* approach, which takes advantage of 3D hand models. Hand states could be recovered by matching the projected 3D model and observed image features, so that the problem becomes a search problem in a high dimensional space. Different image observations have been studied. Fingertips [7, 15, 16] could be used to construct the correspondences between the model and the images. However, the robustness and accuracy largely depends on the performance of fingertip detection. Line features were employed in [12, 14] to enhance the robustness. An exact hand shape model was built by splines in [6], and hand state recovery could be achieved by minimizing the difference between the silhouettes. A method of combining edge and silhouette observations was reported recently for human body tracking [2].

The other alternative is the *appearance-based* approach, which estimates hand states from images directly after learning the mapping from the image feature space to the hand configuration space. In [17], static hand posture recognition is achieved by mapping image feature space to a discrete space of hand configurations. The mapping is highly nonlinear due to the variation in the hand appearances under different view angles. An appearance-based method was also reported in [13] to recover body postures. However, appearance-based approach generally involves a quite difficult learning problem, and it is not trivial to collect large sets of training data.

Fortunately, human motion is often highly constrained. In the case of the hand, the movements of different joints are not independent. Although the degrees of freedom (DoF), D, for the hand is large, the

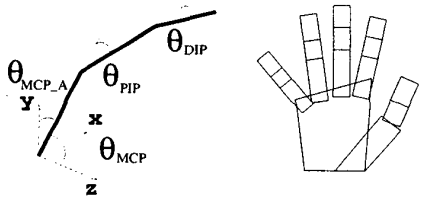
0-7695-1143-0/01 $10.00 © 2001 IEEE

actual hand configuration space could be a small constrained subspace in the state space $\mathcal{R}^D$. The constraints could dramatically reduce the search space in capturing hand articulation. Although some simple and closed form constraints have been found in biometrics and applied to hand motion analysis [7, 6, 16], more investigations on the representations and utilizations of the constraints need to be conducted.

In this paper, we propose an effective method to capture hand articulation by integrating constraints of natural hand motion. Our study of natural hand motion shows that the hand configuration space could be approximated by a lower dimensional space and characterized by a set of basis configurations. To make use of such constraints, an importance sampling based Monte Carlo tracking algorithm is proposed to track hand articulation. Section 2 describes a 3D hand model used in our study. Our study of natural constraints of hand motion is presented in Section 3. Section 4 and Section 5 present the importance sampling technique and our tracking algorithm respectively. Experimental results are shown in Section 6 and we conclude the paper in Section 7.

2 Hand Model

The human hand is highly articulated due to the fact that each finger can be treated as a kinematical chain with palm as its base reference frame. Basically, each finger has four DoFs, two for the MCP joint and one for each of the PIP and DIP joint, as shown in Figure 1(a). The thumb can be approximately modeled by four DoFs. In this sense, hand articulation could be represented by its joint angles $\theta \in \Theta \subset \mathcal{R}^{20}$.



(a) (b)

Figure 1: Hand Model: (a) Kinematical chain of one finger, (b) Cardboard hand model.

When viewed from the direction orthogonal to the palm, the hand could be modeled by a *cardboard* model, in which each finger could be represented by a set of three connected planar patches. The length and width of each patch should be adapted to individual people. The cardboard model is shown in Figure 1(b). Although it is a simplification of the real hand, it offers a good approximation for motion capturing under this specific view direction.

3 Study of Hand Constraints

It is a formidable task to analyze hand articulation in its joint angle space $\Theta \subset \mathcal{R}^{20}$. Fortunately, natural hand articulation is also highly constrained. One type of constraints, usually referred to as static constraints in previous work, are the limits of the range of finger motions as a result of hand anatomy, such as $0^0 \leq \theta_{MCP} \leq 90^0$. These constraints limit hand articulation within a boundary in $\mathcal{R}^{20}$, but without reducing the dimensionality.

Another type of constraints describes the correlations among different joints, and thus reduces the dimensionality of hand articulation. For example, the motions of the DIP joint and PIP joint are generally not independent and they could be described as $\theta_{DIP} = \frac{2}{3}\,\theta_{PIP}$ from the study of biomechanics. Although this constraint could be intentionally made invalid, it is a good approximation of natural finger motion. Unfortunately, not all of such constraints could be quantified in closed forms. These problems motivate us to model the constraints using other alternatives.

Instead of using the joint angle space $\Theta \subset \mathcal{R}^{20}$, we employ hand configuration space Ξ to represent natural hand articulations. We are particularly interested in the dimensionality of the configuration space Ξ and the behaviors of hand articulation in Ξ. To investigate such problems, we propose a learning approach to model hand motion constraints in Ξ using a large set of hand motion data collected using CyberGlove. We have collected a set of more than 30,000 joint angle measurements $\{\theta_k, k = 1, \ldots, N\}$ by performing various natural finger motions. The correlations of different joints are assumed to be well represented by such a data set. The Principal Components Analysis (PCA) technique is employed to project the joint angle space to the configuration space by eliminating the redundancy, i.e.,

$$\mathbf{X} = P^T \cdot (\theta - \theta_0) \tag{1}$$

where P is constructed by the eigenvectors corresponding to large eigenvalues of the covariance matrix of the data set, and $\theta_0 = \frac{1}{N}\sum_{k=1}^{N}\theta_k$ is the mean of the data set. The result shows that we can project the original joint angle space into a 7-dimensional subspace while maintain 95% of information. Thus, $\mathbf{X} \in \Xi \subset \mathcal{R}^7$.

Since the natural hand articulation only covers a subset of $\mathcal{R}^7$, we define 28 basis configurations $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_M : \forall \mathbf{b}_k \in \Xi, M = 28\}$ to characterize the configuration space Ξ. These basis configurations could be identified by clustering the data in Ξ or selecting intuitively. Some of them are shown in Figure 2(a). Sur-

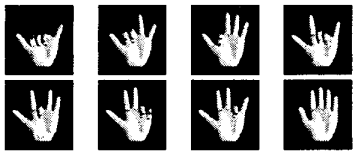
prisingly, after examining the data in Ξ, we found that natural hand articulation lies largely in the linear manifolds spanned by any two basis configurations. For example, if the hand moves from a basis configuration $\mathbf{b}_i$ to another basis $\mathbf{b}_j$, the intermediate hand configuration lies approximately on the linear manifold spanned by $\mathbf{b}_i$ and $\mathbf{b}_j$, i.e.,

$$\mathbf{X} \in \mathcal{L}_{ij} = s\mathbf{b}_i + (1-s)\mathbf{b}_j \tag{2}$$

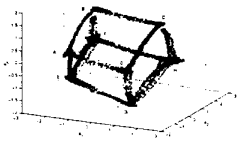
Consequently, hand articulation could be characterized in the configuration space by:

$$\Xi \approx \bigcup_{i,j} \mathcal{L}_{ij}, \text{ where } \mathcal{L}_{ij} = span(\mathbf{b}_i, \mathbf{b}_j) \tag{3}$$

A lower dimensional illustration is shown in Figure 2(b).



(a) a subset of basis configurations



(b) linear manifolds in the configuration space

Figure 2: Hand articulation in the configuration space, which is characterized by a set of basis configurations and linear manifolds.

We noticed that [3] proposed a PCA-based approach to characterize the hand shape deformation, in which hand space deformation lies in the space spanned by a set of eigen shapes. Our representation is different from theirs since our representation characterizes hand articulation in more details. Besides describing a subspace, our representation actually describes the structure of the articulation subset in the configuration space by a union of linear manifolds. Also, our representation of hand articulation is view-independent, since it is derived from the joint angle space.

4 Importance Sampling

A dynamic system could be formulated in a probabilistic framework, and sampling techniques could be used to approximate probabilistic inferences.

4.1 Factored Sampling

In statistics, sampling techniques are widely used to approximate a complex probability density. A set of weighted random samples $\{s^{(n)}, \pi^{(n)}\}, j = 1, \ldots, N$ is *properly weighted* with respect to the distribution f if for any integrable function h,

$$\lim_{N\to\infty} \frac{\sum_{k=1}^{N} h(s^{(k)})\pi^{(k)}}{\sum_{k=1}^{N} \pi^{(k)}} = E_f(h(\mathbf{X}))$$

In this sense, the distribution p is approximated by a set of discrete random samples, $s^{(k)}$ with each having probability proportional to its weight $\pi^{(k)}$.

The tracking problem of a dynamic system could be formulated in a probabilistic framework by representing tracking as a process of conditional probability density propagation. Denote the target state and observation at time t as $\mathbf{X}_t$ and $\mathbf{Z}_t$ respectively, and $\underline{X}_t = \{\mathbf{X}_1, \ldots, \mathbf{X}_t\}, \underline{Z}_t = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_t\}$. The tracking problem is formulated as:

$$p(\mathbf{X}_{t+1}|\underline{Z}_{t+1}) \propto p(\mathbf{Z}_{t+1}|\mathbf{X}_{t+1})p(\mathbf{X}_{t+1}|\underline{Z}_t) \tag{4}$$

Generally, closed-form solutions of dynamic systems are intractable. Monte Carlo methods offer a way to approximate the inference and characterize the evolution of the dynamic systems. Sequential Monte Carlo methods for dynamic systems are studied in the area of statistics [8, 9].

To represent the posteriori $p(\mathbf{X}_t|\underline{Z}_t)$, a set of random samples $\{\mathbf{X}_t^{(n)}, n = 1, \ldots, N\}$ could be drawn from a prior $p(\mathbf{X}_t|\underline{Z}_{t-1})$, and weighted by their measurements, i.e., $\pi_t^{(n)} = p(\mathbf{Z}_t|\mathbf{X}_t = \mathbf{X}_t^{(n)})$, such that the posteriori $p(\mathbf{X}_t|\underline{Z}_t)$ is represented by a set of weighted random samples $\{s_t^{(n)}, \pi_t^{(n)}\}$. This sampling scheme is called *factored sampling* in statistics. It could be shown that such a sample set is properly weighted. This sample set will evolve to a new sample set at time $t+1$ and the new sample set $\{s_{t+1}^{(n)}, \pi_{t+1}^{(n)}\}$ represents the posteriori $p(\mathbf{X}_{t+1}|\underline{Z}_{t+1})$ at time $t+1$. This is the sequential Monte Carlo method employed in CONDENSATION algorithm [4].

CONDENSATION achieved quite robust tracking results. The robustness of Monte Carlo tracking is due to the maintaining of a pool of hypotheses. Since each hypothesis needs to be measured and associated with a likelihood value, the computational cost mainly comes from the image measurement processes. Generally, the more the samples, the more the chances to obtain accurate tracking results but the slower the tracking speed. Consequently, the number of samples becomes an important factor in Monte Carlo based tracking, since it determines the tracking accuracy and speed. Unfortunately, when the dimensionality of the state space

increases, the number of samples increases exponentially.

This phenomenon has been noticed and different methods have been taken to approach this problem by reducing the number of hypotheses. A semi-parametric approach was taken in [1]. It retains only the modes (or peaks) of the probability density and models the local neighborhood surrounding each mode with a Gaussian distribution. Different sampling techniques were also investigated to reduce the number of samples, such as partitioned sampling scheme [10] and annealed particle filtering scheme [2]. [5] emphasized on color-segmented regions by importance sampling.

4.2 Importance Sampling

In practice, it might be difficult to draw random samples from the distribution $f(\mathbf{X})$. Samples could be drawn from another distribution $g(\mathbf{X})$, but their weights should be properly adjusted. This is the basic idea of the technique *importance sampling*. When samples $s^{(n)}$ are drawn from $g(\mathbf{X})$, but weights are compensated as

$$\pi^{(n)} = \frac{f(s^{(n)})}{g(s^{(n)})}\tilde{\pi}^{(n)}$$

It can be proved that the sample set $\{s^{(n)}, \pi^{(n)}\}$ is still *properly weighted* with respect to $f(\mathbf{X})$. This is illustrated in Figure 3.

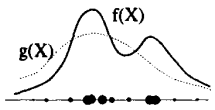


Figure 3: Importance sampling.

To employ the importance sampling technique in dynamic systems, we need to let $f_t(\mathbf{X}_t^{(n)}) = p(\mathbf{X}_t = \mathbf{X}_t^{(n)}|\underline{Z}_{t-1})$, where $f_t(\cdot)$ is the tracking prior, i.e., a prediction density. So, when we want to approximate the posterior $p(\mathbf{X}_t|\underline{Z})$, we could draw random samples from another distribution $g_t(\mathbf{X}_t)$, instead of the prior density $f_t(\mathbf{X}_t)$. But the sample weights should be compensated as

$$\pi_t^{(n)} = \frac{f(\mathbf{X}_t^{(n)})}{g(\mathbf{X}_t^{(n)})}p(\mathbf{Z}_t|\mathbf{X}_t = \mathbf{X}_t^{(n)}) \quad (5)$$

To evaluate $f_t(\mathbf{X}_t)$, we have:

$$\begin{aligned} f_t(\mathbf{X}_t^{(n)}) &= p(\mathbf{X}_t = \mathbf{X}_t^{(n)}|\underline{Z}_{t-1}) \\ &= \sum_{k=1}^{N} \pi_{t-1}^{(k)} p(\mathbf{X}_t = \mathbf{X}_t^{(n)}|\mathbf{X}_{t-1} = \mathbf{X}_{t-1}^{(k)}) \end{aligned}$$

5 Our Approach

This section presents a motion-capturing algorithm based on importance sampling technique. The learned natural hand motion is taken as the alternative tracking prior when using importance sampling technique. Both edge and silhouette are employed as image observation to measure each hypothesis.

5.1 Hypotheses Generating

One important part of sequential Monte Carlo tracking is to generate samples $\{\mathbf{X}_{t+1}^{(n)}, \pi_{t+1}^{(n)}\}$ at time $t+1$ from the samples $\{\mathbf{X}_t^{(n)}, \pi_t^{(n)}\}$ at time t. Instead of directly sampling from the prior $p(\mathbf{X}_{t+1}|\mathbf{Z}_t)$, we propose a method to sample hand articulation manifolds, based on importance sampling technique.

Each hand configuration $\mathbf{X}$ should be either around a basis state $\mathbf{b}_k, k = 1, \ldots, M$, or on the manifold $\mathcal{L}_{ij}$, where $i \neq j, i, j = 1, \ldots, M$. Suppose at time frame t, the hand configuration is $\mathbf{X}_t$. We find the projection $\bar{\mathbf{X}}_t$ of $\mathbf{X}_t$ onto the nearest manifold $\mathcal{L}_{ij}^*$, i.e.,

$$\begin{aligned} \mathcal{L}_{ij}^* &= \arg\min_{i,j} D(\mathbf{X}_t, \mathcal{L}_{ij}) \\ \bar{\mathbf{X}}_t &= Proj(\mathbf{X}_t, \mathcal{L}_{ij}^*) \\ &= \mathbf{b}_i + \frac{(\mathbf{X}_t - \mathbf{b}_i)^T(\mathbf{b}_j - \mathbf{b}_i)}{||(\mathbf{b}_j - \mathbf{b}_i)||}(\mathbf{b}_j - \mathbf{b}_i) \end{aligned}$$

Accordingly,

$$s_t = 1 - \frac{(\mathbf{X}_t - \mathbf{b}_i)^T(\mathbf{b}_j - \mathbf{b}_i)}{||(\mathbf{b}_j - \mathbf{b}_i)||}$$

Then, random samples are drawn from the manifold $\mathcal{L}_{ij}$ according to the density p_{ij}, i.e.,

$$s_{t+1}^{(n)} \sim p_{ij} = N(s_t, \sigma) \quad (6)$$

$$\widetilde{\mathbf{X}}_{t+1}^{(n)} = s_{t+1}^{(n)}\mathbf{b}_i + (1 - s_{t+1}^{(n)})\mathbf{b}_j \quad (7)$$

Then, perform random walk on $\widetilde{\mathbf{X}}_{t+1}^{(n)}$ to obtain hypothesis $\mathbf{X}_{t+1}^{(n)}$, i.e.,

$$\mathbf{X}_{t+1}^{(n)} \sim N(\widetilde{\mathbf{X}}_{t+1}^{(n)}, \Sigma_{t+1}) \quad (8)$$

So, we could write the importance function as: $g_{t+1}(\mathbf{X}_{t+1}^{(n)}) = p(s_{t+1}^{(n)}|s_t)p(\mathbf{X}_{t+1}^{(n)}|\widetilde{\mathbf{X}}_{t+1}^{(n)})$ So,

$$\begin{aligned} g_{t+1}(\mathbf{X}_{t+1}^{(n)}) &\sim \frac{1}{\sigma|\Sigma|^{1/2}} \exp\{-\frac{(s_{t+1}^{(n)} - s_t)^2}{2\sigma^2} \\ &- \frac{1}{2}(\mathbf{X}_{t+1}^{(n)} - \widetilde{\mathbf{X}}_{t+1}^{(n)})\Sigma^{-1}(\mathbf{X}_{t+1}^{(n)} - \widetilde{\mathbf{X}}_{t+1}^{(n)})\} \end{aligned}$$

If the previous hand configuration is one of the basis configurations, say $\mathbf{X}_t = \mathbf{b}_k$, it is reasonable to assume that it takes any one of the manifolds of $\{\mathcal{L}_{kj}, j =$

$1, \ldots, M\}$ with the same probability. Consequently, random samples are drawn from a mixture density p_k:

$$s_{t+1}^{(n)} \sim p_k = \frac{1}{M} \sum_{j=1}^{M} N_{kj}(0, \sigma)$$

Suppose at time t, the tracking posteriori $p(\mathbf{X}_t|\underline{Z}_t)$ is approximated by a set of weighted random samples or hypotheses $\{(\mathbf{X}_t^{(n)}, \pi_t^{(n)}), n = 1, \ldots, N\}$.

In dynamic system, the prior is $p(\mathbf{X}_{t+1}|\underline{Z}_t)$, we have

$$\begin{aligned} f_{t+1}(\mathbf{X}_{t+1}^{(n)}) &= p(\mathbf{X}_{t+1} = \mathbf{X}_{t+1}^{(n)}|\underline{Z}_t) \\ &= \sum_{k=1}^{N} \pi_t^{(k)} p(\mathbf{X}_{t+1} = \mathbf{X}_{t+1}^{(n)}|\mathbf{X}_t = \mathbf{X}_t^{(k)}) \end{aligned}$$

Let

$$p(\mathbf{X}_{t+1}^{(n)}|\mathbf{X}_t^{(k)}) \sim N(A\mathbf{X}_t^{(k)}, \Sigma_2)$$

Instead of sampling directly from the prior $p(\mathbf{X}_{t+1}|\underline{Z}_t)$, samples $s^{(n)}$ could be drawn from another source $g_t(\mathbf{X}_{t+1})$, and the weight of each sample is:

$$\pi_{t+1}^{(n)} = \frac{f_{t+1}(\mathbf{X}_{t+1}^{(n)})}{g_{t+1}(\mathbf{X}_{t+1}^{(n)})} p(\mathbf{Z}_{t+1}|\mathbf{X}_{t+1} = \mathbf{X}_{t+1}^{(n)}) \quad (9)$$

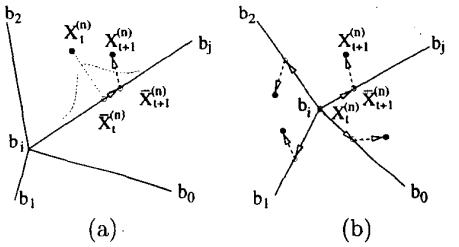


Figure 4: Generating Hypotheses: (a) $\mathbf{X}_t^{(n)} \neq \mathbf{b}_i$, (b) $\mathbf{X}_t^{(n)} = \mathbf{b}_i$.

5.2 Observation Model

We employ both edge and silhouette observations to measure the likelihood of hypotheses, i.e., $p(\mathbf{Z}_t|\mathbf{X}_t)$. Self-occlusion is handled by constructing an occlusion map for the hand model. Since hand is modeled by a cardboard model, it is expected to observe two edges for each planar patch. The cardboard model is sampled at a set of K points on the laterals of the patches. For each such sample, edge detection is performed on the points along the normal of this sample. When we assume that M edge points $\{z_m, m = 1, \ldots, M\}$ are observed, and the clutter is a Poisson process with density λ, then,

$$p_k^e(\mathbf{z}|x_k) \propto 1 + \frac{1}{\sqrt{2\pi}\sigma_e q \lambda} \sum_{m=1}^{M} \exp - \frac{(z_m - x_k)^2}{2\sigma_e^2}$$

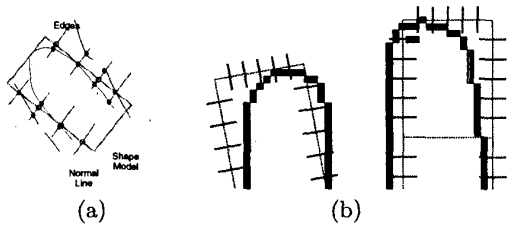


Figure 5: Shape measurements.

We also consider the silhouette measurements, by calculating the difference between the areas of the image A_I and the projected cardboard model A_M, i.e.,

$$p^a \propto \exp - \frac{(A_I - A_M)^2}{2\sigma_a^2}$$

Thus, the likelihood could be written as:

$$p(\mathbf{Z}|\mathbf{X}) \propto p^a \prod_{k=1}^{K} p_k^e \quad (10)$$

5.3 Algorithm Summery

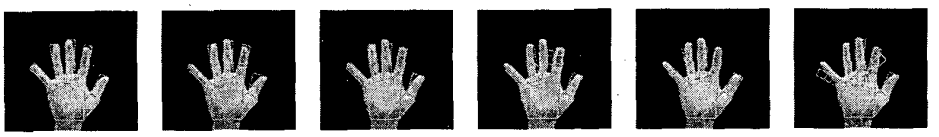
Since natural finger motion could be represented by a set of manifolds in a lower dimensional configuration space, our motion-capturing algorithm takes into account of such motion constraints by importance sampling technique. The motion capturing algorithm is summarized in Figure 6.

6 Experiments

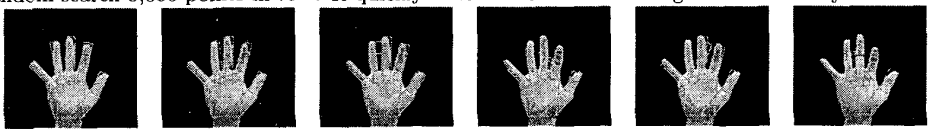
In our experiments, we assume the hand has very little global motion. Only translations in a small range are allowed. Consequentially, the hand motion is represented by $(\mathbf{d}_t, \mathbf{X}_t)$, where $\mathbf{d}_t$ is global 2D translation and $\mathbf{X}_t$ is hand articulation.

We have compared three different methods for both joint angle space $\mathcal{R}^{20}$ and the configurations space $\Xi \subset \mathcal{R}^7$. The first is a random search algorithm, which generates articulation hypotheses based on previous estimate according to a fixed Gaussian distribution without considering any constraints in the joint angle space, i.e., $\theta_{t+1}^{(n)} \sim N(\bar{\theta}_t, \Sigma_\theta)$. The second method is CONDENSATION. The third one is our proposed method based on learned hand constraints using importance sampling.

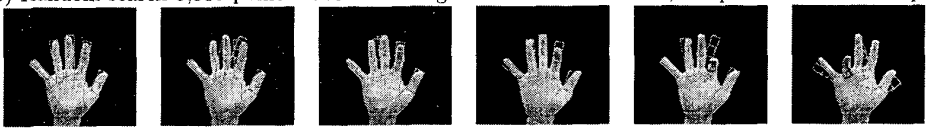
Some experimental results are shown in Figure 6. Figure 6(a) shows the results of random search in $\mathcal{R}^{20}$. We treat each dimension independently with standard deviation of 5^o, and produce 5,000 hypotheses at each frame. However, it hardly succeeded due to the high dimensionality. When we perform random search in



(a) Random search 5,000 points in $\mathcal{R}^{20}$. It quickly losses track due to the high dimensionality of search space.



(b) Random search 5,000 points in $\mathcal{R}^{7}$. Although dimension is reduced, the performance is still poor.



(c) CONDENSATION with 5,000 samples in $\mathcal{R}^{20}$. It does not work well due to the high dimensionality of search space.

(d) CONDENSATION with 3,000 samples in $\mathcal{R}^{7}$. It works fairly well without considering natural motion constraints.

(e) Our approach with 100 samples. Based on the natural motion model, it can track hand articulations in a long sequence.

Figure 7: Comparison of different methods. The projections of the hand model are drawn on the images. When the fingers bend and their backsides appear, the corresponding pieces are drawn in red, otherwise in green.

the reduced space $\mathcal{R}^{7}$ and again with 5,000 hypotheses, the tracking is lost after several frames. The results are shown in Figure 6(b). Figure 6(c) shows some frames of the CONDENSATION in $\mathcal{R}^{20}$, in which 5,000 samples are used. The results show that it is still difficult to handle such a high dimensionality. When performing CONDENSATION in the reduced space $\mathcal{R}^{7}$, the algorithm can track up to 200 frames using 3,000 samples, which is shown in Figure 6(d). Finally, in our proposed algorithm, we use 100 samples, and the algorithm is able to track hand articulations all the time, which is shown in Figure 6(e) [1].

We noticed that our proposed algorithm is efficient, since it is able to perform successful tracking with smaller number of samples compared to CONDENSATION. The reason behind it is that the hand articulation manifolds provide a good prior for tracking, which largely reduces the search complexity.

[1] The demo sequences of our algorithm could be obtained from `http://www.ifp.uiuc.edu/~yingwu`

7 Conclusions

Vision-based capturing of hand articulation is a challenging problem, due to the high degrees of freedom of finger motions. Fortunately, finger movements are also highly constrained, which could be used to ease the high dimensionality problem. In this paper, instead of using the joint angle space, we represented hand articulations in a lower dimensional configuration space, in which hand articulations could be characterized by a set of linear manifolds constructed from 28 basis configurations. Such a representation gives a good approximation of hand articulations. Taking advantage of such a representation of hand articulation, we proposed a sequential Monte Carlo tracking algorithm based on importance sampling technique. The articulation manifolds provide another source of prior to tracking. Our experiments show that this proposed algorithm could perform successful tracking in long sequences efficiently.

Our current capturing algorithm is view-dependent, and the hand model and the method of testing hy-

Monte Carlo Tracking: Probability density propagating from $\{(\mathbf{X}_t^{(n)}, \pi_t^{(n)})\}$ to $\{(\mathbf{X}_{t+1}^{(n)}, \pi_{t+1}^{(n)})\}$, based on importance sampling.

for $n = 1 : N$

// Step(1): Selecting a manifold

if $\mathbf{X}_t^{(n)} \neq \mathbf{b}_k, k = 1, \ldots, M$

$\mathcal{L}_{ij}^* = \arg\min_{i,j} D(\mathbf{X}_t^{(n)}, \mathcal{L}_{ij})$;

$s_t^{(n)} = 1 - \frac{(\mathbf{X}_t^{(n)} - \mathbf{b}_i)^T (\mathbf{b}_j - \mathbf{b}_i)}{\|\mathbf{b}_j - \mathbf{b}_i\|}$;

else

randomly_pick $\mathcal{L}_{ij}^*$;

$s_t^{(n)} = 0$;

// Step(2): Random sampling from $g_t(\cdot)$

$s_{t+1}^{(n)} \sim N(s_t^{(n)}, \sigma)$;

$\widetilde{\mathbf{X}}_{t+1}^{(n)} = s_{t+1}^{(n)} \mathbf{b}_i + (1 - s_{t+1}^{(n)}) \mathbf{b}_j$;

// Step(3): Drifting and diffusing

$\mathbf{X}_{t+1}^{(n)} \sim N(A\widetilde{\mathbf{X}}_{t+1}^{(n)}, \Sigma_1)$;

// Step(4): Observing

$\tilde{\pi}_{t+1}^{(n)} = p(\mathbf{Z}_{t+1} | \mathbf{X}_{t+1} = \mathbf{X}_{t+1}^{(n)})$;

// step(5): Correcting the weights

calculate $f(\mathbf{X}_{t+1}^{(n)})$;

calculate $g(\mathbf{X}_{t+1}^{(n)})$;

$\pi_{t+1}^{(n)} = \frac{f(\mathbf{X}_{t+1}^{(n)})}{g(\mathbf{X}_{t+1}^{(n)})} \tilde{\pi}_{t+1}^{(n)}$;

end

Figure 6: Pseudo code of the sequential Monte Carlo based tracking algorithm.

pothesis are valid only for the view orthogonal to the palm. Some of the hand global motions, such as rotation and scaling, are not considered in our current experiments. Besides the increase of dimensionality, hand global motions would bring about a large amount of self-occlusion. Better methods for testing hypotheses and capturing algorithms including global hand motion will be investigated in our future work.

Acknowledgments

This work was supported in part by National Science Foundation Grants CDA-96-24396 and IRI-96-34618 and NSF Alliance Program. The authors would like to appreciate the anonymous reviewers for their comments.

References

[1] Tat-Jen Cham and James Rehg. A multiple hypothesis approach to figure tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 239–244, 1999.

[2] Jon Deutscher, Andrew Blake, and Ian Reid. Articulated body motion capture by annealed particle filtering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume II, pages 126–133, Hilton Head Island, South Carolina, 2000.

[3] T. Heap and D. Hogg. Towards 3D hand tracking using a deformable model. In *Proc. of IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pages 140–145, Killington, VT, 1996.

[4] Michael Isard and Andrew Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. of European Conf. on Computer Vision*, pages 343–356, Cambridge, UK, 1996.

[5] Michael Isard and Andrew Blake. ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In *Proc. of European Conf. on Computer Vision*, volume 1, pages 767–781, 1998.

[6] James J. Kuch and Thomas S. Huang. Vision-based hand modeling and tracking for virtual teleconferencing and telecollaboration. In *Proc. of IEEE Int'l Conf. on Computer Vision*, pages 666–671, Cambridge, MA, June 1995.

[7] J. Lee and T. Kunii. Model-based analysis of hand posture. *IEEE Computer Graphics and Applications*, 15:77–86, Sept. 1995.

[8] Jun Liu and Rong Chen. Sequential monte carlo methods for dynamic systems. *J. Amer. Statist. Assoc.*, 93:1032–1044, 1998.

[9] Jun Liu, Rong Chen, and Tanya Logvinenko. A theoretical framework for sequential importance sampling and resampling. In *Sequential Monte Carlo in Practice*. 2000.

[10] John MacCormick and Michael Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *Proc. of European Conf. on Computer Vision*, volume 2, pages 3–19, 2000.

[11] Vladimir Pavlović, R. Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human computer interaction: A review. *IEEE PAMI*, 19:677–695, July 1997.

[12] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. of IEEE Int'l Conf. Computer Vision*, pages 612–617, 1995.

[13] Romer Rosales and Stan Sclaroff. Inferring body pose without tracking body parts. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 721–727, 2000.

[14] Jakub Segen and Senthil Kumar. Shadow gesture: 3d hand pose estimation using a single camera. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 479–485, 1999.

[15] N. Shimada and et al. Hand gesture estimation and model refinement using monocular camera - ambiguity limitation by inequality constraints. In *Proc. of the 3rd Conf. on Face and Gesture Recognition*, pages 268–273, 1998.

[16] Ying Wu and Thomas S. Huang. Capturing articulated human hand motion: A divide-and-conquer approach. In *Proc. IEEE Int'l Conf. on Computer Vision*, pages 606–611, Corfu, Greece, Sept. 1999.

[17] Ying Wu and Thomas S. Huang. View-independent recognition of hand postures. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume II, pages 88–94, Hilton Head Island, South Carolina, June 2000.