3D MODEL-BASED VISUAL HAND TRACKING

Thomas S. Huang[†], *Ying Wu*[‡], *John Lin*[†]

[†] University of Illinois at Urbana-Champaign, 405 N. Mathews, Urbana, IL 61801
[‡] Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3118
{huang, jy-lin}@ifp.uiuc.edu, yingwu@ece.nwu.edu

ABSTRACT

Capturing human hand motion through visual input is a challenging problem that involves the estimation of both global hand pose as well as the local finger articulation. This is a difficult task that requires a search in a high dimensional space due to the high degrees of freedom that fingers exhibit and the self occlusions caused by global hand motion. In this paper we propose a divide and conquer approach to estimate both global and local hand motion. The hand pose is determined from the palm using Iterative Closed Point (ICP) algorithm and factorization method. By incorporating natural hand motion constraints, we propose an efficient tracking algorithm based on sequential Monte Carlo technique for tracking finger motion. Finally, the iteration step between the pose estimation and finger articulation tracking is performed in an EM fashion to obtain an accurate configuration estimation. Our experiments show that our approach is accurate and robust for natural hand movements.

1. INTRODUCTION

Rather than using the mouse and the keyboard, hand gestures can be used as a more natural and convenient way for human to communicate with computers. Several applications such as the virtual environment interaction, would benefit directly from such an interface. One important component for this interface is how to capture hand motions. As an alternative to glove-based techniques, vision-based techniques offer an non-intrusive and affordable approach to hand motion capturing. However, this task is difficult due to the high degrees of freedoms involved.

Different methods have been proposed to analyze human hand motion for visual hand tracking. One choice is the appearance-based approach, which tries to establish the mapping between the image feature space and the hand motion space [1, 2]. However, the mapping can be difficult to learn and may not be one-to-one. Also, it is not trivial to collect large and representative set of training data.

Another approach is the 3D model-based approach. The hand motion could be estimated by matching the 3D model projections and observed image features, so that the problem becomes a search problem in a high dimensional space. To construct the correspondences between the model and the images, different image observations have been studied, such as fingertips [3, 4, 5], line features [6], and silhouettes [7, 8, 9].

Many methods tackle the global hand motion and local finger motion simultaneously, such that the optimization would have a very high chance to converge to a local minima. On the other hand, a divide-and-conquer approach [5] could be taken to separate the hand pose determination and articulate estimation.

This paper proposes a model-based approach to capture both hand pose and finger articulation in the divide-and-conquer framework. In Section 2 we describe how to represent the hand model and hand motion constraint. Section 3 presents an approach to combine global and local motion. In Section 4 a method is given for estimating global motion using ICP. Section 5 shows an algorithm for estimating local motion. Our experiment results are presented in Section 6. Finally, conclusions are given in Section 7.

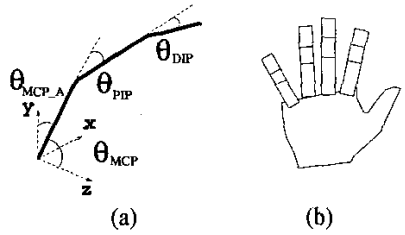
2. HAND MODEL AND MOTION CONSTRAINTS

The hand motion consists of global hand pose M_G and local finger articulation M_L. Global hand motion can be described by 3D translation t and rotation $\mathbf{R}$ of the palm. The local finger motion is represented by a set of joint angle Θ. The hand structure is represented by a kinematical model (Figure 1a), which has roughly 20 degrees of freedom [3, 7]. The task of motion capturing is to estimate $\{\mathbf{R}, t, \Theta\}$.

In our experiment, we use a cardboard model in which each finger is represented by a set of three connected planar patches. The parameters of the patches are calibrated according to individual user. (Figure 1b). Although it is a simplification of the human hand structure, it offers a good approximation for motion capturing.

Instead of searching in the 20 dimensional space, we would like to use various constraints to reduce the dimensionality of the joint angle space and find a smaller feasible space, which we will call the configuration space Ξ. Sev-

0-7803-7304-9/02/$17.00 C2002 IEEE



(a) (b)

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

eral commonly known constraints due to the anatomy of the hand can be used to initially reduce the dimensionality to roughly 15. To further reduce the dimensionality, we have collected more than 30,000 joint angle data from various hand motions using CyberGlove. Then PCA is applied to eliminate the redundancy. We can project $\Theta \in \mathcal{R}^{20}$ into a 7-dimensional subspace while maintaining 95% of the information. Therefore, the configuration space Ξ is defined in $\mathcal{R}^7$. Furthermore, we define 28 basis configurations as follows. For each basis state $\mathbf{b}_i$, each finger is either fully extended or fully curled. Our observations of the motion trajectories between basis states in Ξ show that they are roughly linear and that natural hand articulation can be characterized by these linear manifold $\mathcal{L}_{ij}$ spanned by $\mathbf{b}_i$ and $\mathbf{b}_j$, with $i \neq j$.

$$\Xi \approx \bigcup_{i,j} \mathcal{L}_{ij}, \text{ where } \mathcal{L}_{ij} = span(\mathbf{b}_i, \mathbf{b}_j) \qquad (1)$$

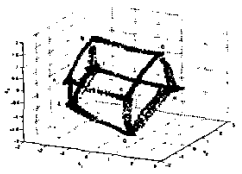


Figure 2: Hand articulation in the configuration space, which is characterized by a set of basis configurations and linear manifolds.

3. DIVIDE AND CONQUER APPROACH

Rather than estimating both global and local motion altogether, another approach is to estimate global and local motions separately and combine the results in an iterative manner [5]. The idea is to use additional information from finger projections and iterate between two steps: (1) pose determination based on palm contour and some extra points, using the method describe in Section 4; (2) tracking local finger configurations using a Monte Carlo based algorithm as described in Section 5. The iterations between global and local hand motion estimation would converge to a local stationary point that minimizes the discrepancies between the image observation and model projection.

4. CAPTURING GLOBAL MOTION

The global hand motion is defined by the pose of the palm, which is treated as a rigid planar object. In this section, we present algorithms for determining the pose and estimating the global motion.

4.1. Iterative Closed Points

We first describe a method for establishing point correspondences by adapting the idea of the Iterative Closed Point (ICP) algorithm [10]. The basic idea is to refine the correspondences and the motion parameters iteratively.

The ICP algorithm takes the image edge point that is closest to the projected 3D model point as its correspondence. Motion parameters $\{\mathbf{R}, t\}$ can be computed based on these temporary correspondences using the pose determination method presented in Section 4.2. The computed motion would result in a new matching. Iteratively applying this procedure, ICP would continue to yield an improved pose estimation. It should be pointed out that ICP procedure converges only to local minima, which means that we need a close initial start.

4.2. Pose Estimation

After the correspondences have been constructed, we may determine the pose using the following approach. Let a point on the plane be $\mathbf{x}_i = [x_i, y_i]^T$, and its image point be $\mathbf{m}_i = [u_i, v_i]^T$. Under scaled orthographic projection, we can write $t_3\mathbf{m}_i = \mathbf{A}\mathbf{x}_i + \mathbf{t}$, where

$$\mathbf{A} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \text{ and } \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

We can subtract the centroid of the projection points and model points, i.e., $\hat{\mathbf{m}}_i = \mathbf{m}_i - \bar{\mathbf{m}}$ and $\hat{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, which gives $t_3\hat{\mathbf{m}}_i = \mathbf{A}\hat{\mathbf{x}}_i$. If we let $\mathbf{B} = \mathbf{A}/t_3$, then we have $\hat{\mathbf{m}}_i = \mathbf{B}\hat{\mathbf{x}}_i$

Denoting $[u_i^k, v_i^k]^T$ to be the i-th image point at the k-th frame, we can write

$$\mathbf{W} = \begin{bmatrix} u_1^1 & u_2^1 & \dots & u_N^1 \\ v_1^1 & v_2^1 & \dots & v_N^1 \\ u_1^2 & u_2^2 & \dots & u_N^2 \\ v_1^2 & v_2^2 & \dots & v_N^2 \end{bmatrix} = \mathbf{MS}$$

where

$$\mathbf{M} = \begin{bmatrix} M^1 \\ M^2 \end{bmatrix} \text{ and } \mathbf{S} = \begin{bmatrix} x_1 & x_2 & \dots & x_N \\ y_1 & y_2 & \dots & y_N \end{bmatrix}$$

The factorization method [11] can be used to solve for $\mathbf{M}$ and $\mathbf{S}$ up to a matrix $\mathbf{D}$, which could be determined by the constraints of $\mathbf{M}$.

After recovering $\mathbf{M}$, which contains $\mathbf{B}$ and t_3, $\mathbf{R}$ and t_3 may be computed from $\mathbf{B}$. Once the rotation matrix $\mathbf{R}$ and the depth translation t_3 are computed, we can compute:

$$\begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = t_3\bar{\mathbf{m}} - \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \bar{\mathbf{x}}$$

For simplicity, we can use the first frame that shows the front of the palm for initialization and calibration, and take image points along the palm contour as the model points.

5. CAPTURING FINGER ARTICULATION

In this section, we present a sequential Monte Carlo algorithm that takes advantage of the natural hand motion constraints in the tracking algorithm.

5.1. Sequential Monte Carlo

The tracking problem could be formulated as a process of conditional probability density propagation. We can track the finger motions efficiently using sequential Monte Carlo method, which offers a way to approximate the evolution of the densities. Denote the target state and image observations by $\mathbf{X}_t$ and $\mathbf{Z}_t$ respectively, and $\underline{Z}_t = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_t\}$, the tracking problem is formulated as:

$$p(\mathbf{X}_{t+1}|\underline{Z}_{t+1}) \propto p(\mathbf{Z}_{t+1}|\mathbf{X}_{t+1})p(\mathbf{X}_{t+1}|\underline{Z}_t) \quad (2)$$

The posteriori $p(\mathbf{X}_t|\underline{Z}_t)$ can be represented by a set of random samples $\{s_t^{(n)}, \pi_t^{(n)}\}$ which will evolve to a new set of samples $\{s_{t+1}^{(n)}, \pi_{t+1}^{(n)}\}$ at time $t+1$ to represent the new posteriori. Different sampling schemes can be used depending on the source of sampling priors [9, 12].

Since finger articulation involves a high DOF, algorithms such as CONDENSATION will require a large number of samples for representing the density propagation, and an intensive computation will be unavoidable. Fortunately, we may reduce the complexity by making use of the finger motion constraints as an outside prior for the importance sampling technique. Let $f_t(\mathbf{X}_t^{(n)}) = p(\mathbf{X}_t = \mathbf{X}_t^{(n)}|\underline{Z}_{t-1})$, be the tracking prior. 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 $f_t(\mathbf{X}_t)$. Below we will give a brief description of this method. The details can be found in [9].

For natural hand motion, 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 $\mathcal{L}_{ij}^*$, and obtain

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

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

$$s_{t+1}^{(n)} \sim p_{ij} = N(s_t, \sigma) \quad (3)$$

$$\widetilde{\mathbf{X}}_{t+1}^{(n)} = s_{t+1}^{(n)}\mathbf{b}_i + (1 - s_{t+1}^{(n)})\mathbf{b}_j \quad (4)$$

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

$$\mathbf{X}_{t+1}^{(n)} \sim N(\widetilde{\mathbf{X}}_{t+1}^{(n)}, \Sigma_{t+1}) \quad (5)$$

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,

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

Finally, the weights must be properly compensated:

$$\pi_{t+1}^{(n)} = \frac{f_{t+1}(\mathbf{X}_{t+1}^{(n)})}{g_{t+1}(\mathbf{X}_{t+1}^{(n)})} p(\mathbf{Z}_{t+1}|\mathbf{X}_{t+1} = \mathbf{X}_{t+1}^{(n)}) \quad (6)$$

If the previous hand configuration is at one of the basis configurations, say $\mathbf{X}_t = \mathbf{b}_k$, it is reasonable to assume that it selects 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.

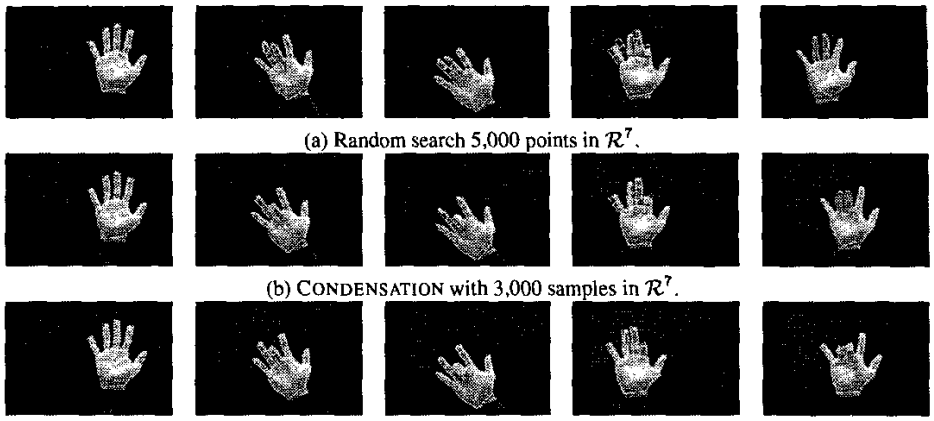
5.2. Model Matching

We employ edge observations to measure the likelihood of hypotheses, i.e., $p(\mathbf{Z}_t|\mathbf{X}_t)$ as in [9]. Self-occlusion is handled by constructing an occlusion map for the hand model. The cardboard model for the hand is sampled at a set of K points on the laterals of the patches. For each of these samples, 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,

$$p(\mathbf{Z}|\mathbf{X}) \propto \prod_{k=1}^{K} \left(1 + \frac{1}{\sqrt{2\pi}\sigma_e q \lambda} \sum_{m=1}^{M} \exp -\frac{(z_m - x_k)^2}{2\sigma_e^2}\right)$$

6. EXPERIMENT

We have tested our algorithm on real hand motion sequences. Different schemes are also compared for local motion capturing. The first one is a random search scheme in the $\mathcal{R}^7$



(a) Random search 5,000 points in $\mathcal{R}^7$.

(b) CONDENSATION with 3,000 samples in $\mathcal{R}^7$.

(c) Our approach with 100 samples.

Figure 3: *Comparisons of different methods on real sequences. Our method is more accurate and robust than the other two methods.*

space. We use 5000 random samples, but since it makes use of no constraints, the performance is poor for local motion estimation and also degrades the global pose estimation. The second scheme uses CONDENSATION with 3000 samples in $\mathcal{R}^7$. It performs better than the first method, but it is still not robust enough. The third scheme is the proposed method, and it works accurately and robustly. The articulation model makes the computation more efficient and the local motion estimation enhances the accuracy of hand pose determination.

7. CONCLUSIONS

Recovering hand motions from video sequences is a difficult problem due to the high degrees of freedom involved. This paper presents a divide and conquer approach to this problem by separating the global and local hand motion and estimating each component separately. For the global motion, we approximate the palm as a rigid planar object and use ICP to track the hand pose. The local finger articulation is tracked through a sequential Monte Carlo technique. The iterations between the estimates of global and local finger motion result in an accurate motion estimation. There are still several directions for future extensions. For instance, a better hand model could be used to handle out of plane rotations. Also, it would be more interesting if we could achieve automatic initialization for the tracking.

8. ACKNOWLEDGEMENTS

This work was supported in part by National Science Foundation Grants CDA-96-24396 and EIA 99-75019, and NSF Alliance Program.

9. REFERENCES

[1] Romer Rosales, Stan Sclaroff, and Vassilis Athitsos, "3D hand pose reconstruction using specialized mappings," in *Proc. IEEE Int'l Conf. on Computer Vision*, Vancouver, Canada, July 2001.

[2] Ying Wu and Thomas S. Huang, "View-independent recognition of hand postures," *in Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2000, vol. II, pp. 88–94.

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

[4] N. Shimada, Y. Shirai, Y. Kuno, and J. Miura, "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*, 1998, pp. 268–273.

[5] Ying Wu and Thomas S. Huang, "Capturing articulated human hand motion: A divide-and-conquer approach," in *Proc. of IEEE Int'l Conf. Computer Vision*, 1999, pp. 606–611.

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

[7] 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*, Cambridge, MA, June 1995, pp. 666–671.

[8] 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*, Hilton Head Island, South Carolina, 2000, vol. II, pp. 126–133.

[9] Ying Wu, John Lin, and Thomas S. Huang, "Capturing natural hand articulation," in *Proc. of IEEE Int'l Conf. Computer Vision*, 2001, vol. II, pp. 426–432.

[10] Zhengyou Zhang, "Iterative point matching for registration of free-form curves and surfaces," *Int'l Journal of Computer Vision*, vol. 13, pp. 119–152, 1994.

[11] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography - a factorized method," *Int'l Journal of Computer Vision*, vol. 9, pp. 137–154, 1992.

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