Articulate Hand Motion Capturing Based on a Monte Carlo Nelder-Mead Simplex Tracker

John Lin†, Ying Wu‡, Thomas S. Huang†

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

Abstract

This paper presents an algorithm for tracking the articulate hand motion in monocular video sequences. The task is challenging due to the high degrees of freedom involved in the hand motion. The complexity can be reduced by considering the natural motion constraints. To take advantage of the constraints, we propose to use a nonparametric representation of the feasible configuration space and employ a Monte Carlo Nelder-Mead simplex search algorithm. The tracker combines the strengths of both sequential Monte Carlo and direct search algorithms. First, its multiple hypotheses nature increases the chance of the simplex method to identify the global maximum. Second, the direct search algorithm produces a set of more representative particles. Experiment results show that this hybrid approach is robust for tracking the hand motion.

1. Introduction

One of the fundamental component in a natural HCI interface is the human motion understanding from visual input. Although the most reliable results are still obtained from the use of external sensors, recent advances demonstrate that vision-based techniques can offer an inexpensive and non-invasive alternative for capturing human motions. In this paper we present a model-based approach for tracking the highly articulate human hand motion. One of the main challenge is the large degrees of freedom (DOF) involved in the human hand motion; thus, estimating the correct hand configuration parameters becomes equivalent to a search problem in the high dimensional space which forbids exhaustive or simple search without any prior knowledge.

The model-based hand tracking approach [5, 6, 7, 8, 9, 10, 13] can produce a very accurate estimate when the tracker and model are well initialized. This approach compares real hand image features to several hand model projections. The model configuration that generates the best match determines the current hand state. However, model-based hand tracking requires the estimation of roughly 27 parameters. To cope with the high DOF problem, previous works have shown that the dimensionality of the feasible space can be significantly reduced by considering motion constraints [5, 8, 13]. To incorporate the motion constraints in the model-based approach, we must determine the structure of the feasible configuration space, and employ an efficient search algorithm associated with this representation. Previous works [2, 3] generally identify the lower-dimensional manifold using piecewise linear assumption. A clustering algorithm is first applied to produce locally similar patches, which is then approximated by a linear manifold in a lower dimensional space. It is generally a difficult problem in determining a suitable clustering algorithm and to construct an accurate manifold approximation.

We propose a novel representation for the feasible configuration space using a set of discrete samples collected from CyberGlove. Each sample corresponds to one hand configuration. By using the entire set of samples directly, we avoid the representation error due to incorrect structure assumptions and lossy dimensionality reduction techniques. To search in this discrete space, we propose to use a Monte Carlo Nelder-Mead (NM) simplex search. NM method is a classical direct search algorithm that is used for the cases when gradients can not be accessed or evaluated. The multiple hypotheses variant is employed to effectively deal with the local minima problem when a nonconvex objective function is used.

2. The Feasible Configuration Space

Although the global and local hand motions can be estimated separately, the finger motion still involves roughly 20 degrees of freedom (DOFs) [5]; therefore, exhaustive searching without any prior knowledge of the feasible space

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

is a nearly impossible task. How to model natural configuration distribution of the feasible states is the key in successfully reducing the computational complexity. Wu *et al* [13] showed that it is possible to reduce the dimensionality to 7 using anatomical constraints and PCA.

We propose to adopt a nonparametric representation and model the feasible space directly from the set of N_C collected Cyberglove data. The entire feasible space Ψ is defined by the set $\{\theta_i, i = 1 \dots N_C\}$, where each $\theta \in \mathcal{R}^{20}$ represents one sampled configuration. Then a kd-tree structure is constructed so that given any point $\theta \in \mathcal{R}^{20}$, we can quickly find an approximated nearest neighbor $\theta' \in \Psi$. One of the benefits of using this representation is that no learning is required to find a closed form of the manifold parametrization of Ψ. Therefore, this approach avoids the error induced from incorrect manifold approximations. The model can be better refined when more samples are collected. This advantage is gained as a trade off for the cost of the computer memory, which is inexpensive nowadays.

3. The Direct Search Algorithm

3.1. Nelder-Mead Simplex Algorithm

With the feasible configuration space Ψ defined, we need to design a search algorithm that is appropriate for this structure. Given the object state x_t at time t, the goal is to identify x_{t+1} which minimizes certain objective function $f(x)$. We choose the non-gradient Nelder-Mead (NM) method [11] because Ψ has several properties that make it unfavorable for the numerical optimization. First, because Ψ uses a nonparametric representation in a high dimensional space, it is difficult to obtain an estimate of the derivative for the objective function $f(x)$. Second, although we could define $f(x)$ for every x, it is difficult to obtain the closed form of $f(x)$ due to its nonlinear nature. Third, the representation has several discontinuities due to the existence of many infeasible configurations. Because of these properties, it is impossible to obtain the gradient and we must rule out the gradient descent algorithms, such as Newton methods.

The NM method maintains at each iteration a nondegenerate simplex S, which is a geometric object defined by a convex hull of $n+1$ points $\{x_0, \dots, x_n\}$in n dimensional space. Through a sequence of elementary geometric transformations, the initial simplex moves towards the minimum. At each step, the worst vertex with highest cost $x_{max} = \arg\max_{x \in S} f(x)$ is selected and reflected with respect to the centroid $\overline{x} = \frac{1}{n}(\sum_{i=0}^{n} x_i - x_{max})$ to obtain x_r. Then depending on $f(x_r)$, we either keep x_r, or perform an expansion or a contraction to acquire the new vertex x_{new}which replaces x_{max} such that $f(x_{new}) < f(x_{max})$. The iteration for NM-method typically terminates for a sufficiently small simplex or when a maximum number of iteration is reached.

3.2. Two Stage NM Method

The basic form of NM method does not take advantage of the motion constraints embedded in Ψ. To incorporate the constraints in the search, we implement a two stage hierarchical simplex search. In the coarse search phase, we begin with a larger simplex and restrict the simplex vertices θ_i to be one of the samples $\theta_j \in \Psi$. At the k^{th} iteration, a new vertex θ_{new} is generated as described in Sec. 3.1, and a nearby configuration θ' is located to replace θ_{max} for S_t^{k+1}.

$$\theta_t^{k+1} = \theta' = [\theta_{new}]^+ = \arg\min_{\theta \in \Psi} \parallel \theta_{new} - \theta \parallel$$

By constraining the searching to the discrete space Ψ, the hand motion constraints are automatically enforced in the searching. Since the data we collected can not possibly cover the entire feasible space, there exist gaps and discontinuities in Ψ. Searching only in the discrete domain will not guarantee an optimal convergence; therefore, after the initial simplex converges to a smaller region in the first stage, we must continue the iteration in the continuous domain with a more strict termination condition.

4. Monte Carlo NM Simplex Tracking

The NM simplex search fails when the objective function contains several local minima. Because of the noise presented in the image feature extraction, and the nontrivial definition of the cost function (Sec 5), the objective function can not be convex. One approach to tackle this problem is to utilize the multiple hypotheses approach and run several simplex searches at each frame. A well established formulation of the probabilistic tracking algorithm that uses multiple hypotheses can be shown with the Bayes rule [1]:

$$p(x_{t+1}|\underline{\mathbf{z}}_{t+1}) \propto p(\mathbf{z}_{t+1}|x_{t+1}) p(x_{t+1}|\underline{\mathbf{z}}_t) \qquad (1)$$

where x_t is the target state at time t and $\underline{\mathbf{z}}_t = \{\mathbf{z}_1, \dots \mathbf{z}_t\}$ is the history of image observations. A practical implementation of the algorithm is the sequential Monte Carlo simulation aka particle filtering, which uses a set of N random samples $\{s_t^{(n)}, \pi_t^{(n)}\}$ to approximate arbitrary nonlinear multi-modal pdf. One of the problem with particle filtering is the degeneracy phenomenon where the weights for many of the samples become insignificant during the evolution process and a large computational effort is wasted to maintain these samples. Many algorithms were suggested to reduce this effect, such as resampling [4] and importance sampling [13].

We propose to combine both NM method and particle filtering in order to take advantage of both approaches. Instead of using a set of random samples to model the pdf

evolution, we use a set of simplices each generated from a mode of the pdf. The new algorithm shares the advantages of each approach to reduce the limitations induced by employing NM simplex search or SMC alone. First, the implementation of multiple hypotheses increases the chances of reaching the global minimum. Second, the prior $p(x_{t+1}|\underline{\mathbf{z}}_t)$ (Eq. 1) is considered in estimating a more accurate hand state estimation. Third, each of the new sample generated will be close to a mode with a significant weight. Although forcing every sample to be at one of the peak will lead to sample impoverishment, we argue that the procedure for generating the initial simplex from a point is similar to a perturbation procedure which will increase the diversity of sample representation.

The algorithm begins by drawing random samples $\tilde{x}_t^i$, $i = 1, \ldots, N$ from $\{s_t^{(n)}, \pi_t^{(n)}\}$ based on $p(x_t|\underline{\mathbf{z}}_t)$. Then an initial simplex S^{i0} is generated from each $\tilde{x}_t^i$, which corresponds to a mode in $p(x_t|\underline{\mathbf{z}}_t)$. To incorporate the linear manifold constraint observed by Wu [13], each of the vertex in the initial simplex is generated from the importance function $q(x_t|\underline{\mathbf{z}}_t)$ as described in [13]. Next the two stage simplex search is carried out to obtain a local minimum corresponding to a mode of the pdf:

$$S^{i*\prime} = \mathcal{NM}^d(S^{i0})$$

$$S^{i*} = \mathcal{NM}^c(S^{i*\prime}) \tag{2}$$

where $\mathcal{NM}^d$ and $\mathcal{NM}^c$ denotes the discrete and continuous NM operations respectively. Each operation takes an initial set of vertices and outputs a converged simplex. The search terminates when

$$\sum_{j=0}^{n} \parallel x_j^k - x_j^{k+1} \parallel^2 < \epsilon \tag{3}$$

where $x_j^k \in S^k$ and S^k is the simplex generated at the k^{th} iteration. The new sample x_{t+1}^i is the centroid of the converged simplex:

$$x_{t+1}^i = \frac{1}{n+1} \sum_{x_j \in S^{i*}} x_j \tag{4}$$

If the simplex converged according to Eq. 3, then the weight $\pi_{t+1}^i = p(\mathbf{z}_{t+1}|x_{t+1})$ is computed as described in Sec. 5. Otherwise $\pi_{t+1}^i = 0$.

5. Experiments

In our experiments, we use a 3D hand model with each finger phalanx represented using a truncated cylinder. The continuous space stochastic NM simplex search is applied for global parameter estimation. Then the two stage simplex search (Sec. 4) is employed to recover the finger motion. These two steps are repeated until the results converge [12]. To measure the likelihood of hypothesis, the hand model is first projected onto the image plane as described in [10] to obtain the edge points that define the projected shape. If K projected edge samples are generated, edge detection is performed on the points along the normal of this sample. Assuming that M image edge points $\{z_m, m = 1, \ldots, M\}$ are observed, and the clutter is a Poisson process with density λ, then,

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

We noticed that with edge points alone could not provide a good likelihood estimation. Therefore we also consider the silhouette measurement. The segmented foreground pixels are XORed with the model silhouette image, and the likelihood is computed as $p^s \propto \exp - \frac{(A_I - A_M)^2}{2\sigma_s^2}$. Since a well matched projection contributes lower cost, the objective function at time t is defined as the negative of the likelihood function:

$$f(x, \mathbf{z}) = -p(\mathbf{z}|x) \propto -p^s \prod_{k=1}^{K} p_k^e \tag{5}$$

In the video sequence, the fingers bend and extend while the hand moves simultaneously (Figure 1). We used 30 simplices for finger articulation tracking and 10 simplices for global motion. The algorithm takes about 2 sec/frame to run on an Intel 2GHz PC. We have also tested the sequence using CONDENSATION algorithm with 5000 samples, and the algorithm fails after about 10 frames. The projection of each estimated configuration is superimposed on the hand image, and a reconstructed 3D hand model is shown below each corresponding image for better visualizations. The experiment results show that our algorithm is robust and successful in tracking complex hand motions in a cluttered environment.

6. Conclusions

This paper proposes to track the articulate hand motion by constructing a nonparametric representation of the feasible configuration space and employing a multiple hypotheses variant of NM simplex search algorithm. The feasible space is modelled directly from real hand motion data and avoids the errors induced from incorrect manifold assumptions. The NM simplex search algorithm, which requires no knowledge of gradients, is particularly suitable for searching in this discrete space. Since direct search methods are often trapped in local optima, we extend the algorithm to embed the NM search in a particle filter. The experiment results show that our algorithm is robust in tracking the hand

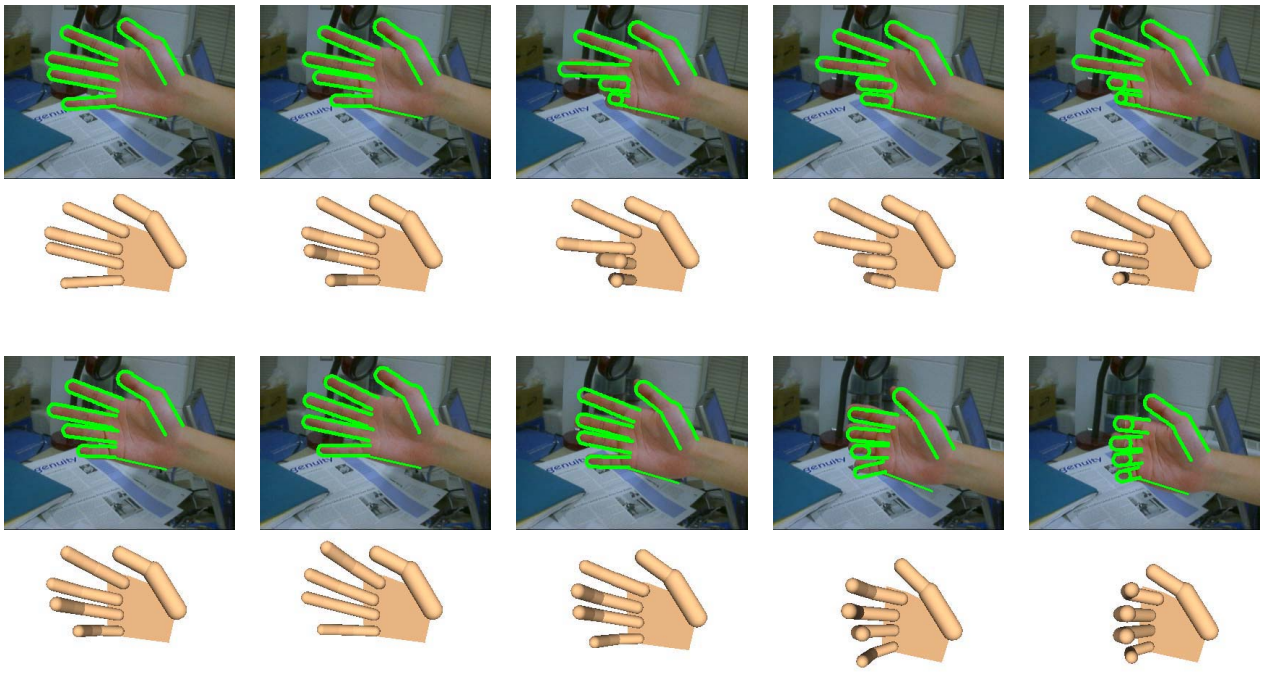

Figure 1. Simultaneously tracking finger articulation and global hand motion. The projected edge points are superimposed with the real hand image. Below each real hand image, a corresponding reconstructed 3d hand model is shown for better visualization.

motions in cluttered background. We are currently studying the convergence analysis of this algorithm and the extension to general tracking problems. Another future extension of the work would be to incorporate temporal constraints in the tracker.

Acknowledgments

The authors thank Howard Huang for his numerous insightful comments and Arjun Kulothungun for his contribution on model projection. This work was supported in part by National Science Foundation Grant IIS-01-38965, and NSF IIS-0347877 for YW.

References

[1] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions of Signal Processing*, 50(2):174–188, 2002.

[2] R. Basri, D. Roth, and D. Jacobs. Clustering appearances of 3d objects. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 414–420, 1998.

[3] A. Heap and D. Hogg. Wormholes in shape space: Tracking through discontinuous changes in shape. In *Proc. of IEEE Int'l Conf. Computer Vision*, pages 344–349, 1998.

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

[5] J. J. Kuch and T. 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.

[6] S. Lu, D. Metaxas, D. Samaras, and J. Oliensis. Using multiple cues for hand tracking and model refinement. In *Proc. IEEE Int'l Conf. on Computer Vision*, volume II, pages 443–450, Madison, 2003.

[7] V. Pavlović, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human computer interaction: A review. *IEEE Trans. on PAMI*, 19:677–695, July 1997.

[8] 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.

[9] 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*, pages 268–273, 1998.

[10] B. Stenger, P. R. S. Mendonça, and R. Cippola. Model-based 3d tracking of an articulated hand. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Hawaii, 2001.

[11] F. Walters, L. R. Parker, S. L. Morgan, and S. N. Deming. *Sequential Simplex Optimization*. CRC Press, Boca Raton, USA, 1991.

[12] Y. Wu and T. S. Huang. Capturing articulated human hand motion: A divide-and-conquer approach. In *Proc. of IEEE Int'l Conf. Computer Vision*, pages 606–611, 1999.

[13] Y. Wu, J. Lin, and T. S. Huang. Capturing natural hand articulation. In *Proc. of IEEE Int'l Conf. Computer Vision*, volume II, pages 426–432, 2001.