3D Model-Based Hand Tracking Using Stochastic Direct Search Method

John Y. 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

Tracking the articulated hand motion in a video sequence is a challenging problem in which the main difficulty arises from the complexity of searching for an optimal motion estimate in a high dimensional configuration space induced by the articulated motion. Considering that the complexities of this problem may be reduced by learning the lower dimensional manifold of the articulation motion in the configuration space, we propose a new representation for the nonlinear manifold of the articulated motion, with a stochastic simplex algorithm that facilitates very efficient search. Contrary to traditional methods of representing the manifolds through clustering and transition matrix construction, we maintain the set of all training samples. To perform the search of best matching configuration with respect to the input image, we combine sequential Monte Carlo technique with the Nelder-Mead simplex search which is efficient and effective when the gradient is not readily accessible. This new approach has been successfully applied to hand tracking and our experiments show the efficiency and robustness of our algorithm.

1. Introduction

Capturing the articulate hand motion from visual input is an important task in many fields with various applications including human motion understanding, gesture recognition, and motion capture for animation synthesis. Currently, the most reliable methods still require external sensors attached to the human body such as the special markers for the motion capture systems, data gloves for hand motion capturing, and magnetic sensors for virtual environment interactions. These devices are generally expensive and cumbersome.

On the other hand, vision-based techniques offer an inexpensive and non-invasive alternative for capturing the articulate body motions. In this paper we present an approach for tracking the highly articulate human hand motion. The difficulties of tracking the human hand arises mainly from the large degrees of freedom (DOF) involved in the motion. As a result, estimating the correct hand motion and configuration parameters is equivalent to a search in the high dimensional space. Other difficulties include the self occlusions of different fingers, and clutters from the image background which introduce additional uncertainties for the estimation.

One strategy for tracking the hand motion is using the appearance-based approach [2, 7, 16, 17, 23], which attempts to estimate the hand states directly from the image features. A nonlinear mapping is learned from a large amount of training images. This approach can quickly estimate the hand configuration once the mapping is learned. However, it is difficult to determine the structure of the mapping function and the set of optimal training data.

Model-based approach is another alternative for estimating hand articulations [10, 11, 13, 14, 15, 18, 20, 24, 19]. The idea is to compare image features between 3D hand model projections and real hand images. The hand state is recovered from the configuration that generates the best match. With a well initialized hand model, this approach can produce a very accurate estimate. However, model-based tracking involves the estimation of parameters in high dimensional space. The hand motion generally consists of 27 DOF, 6 for the global motion and 21 for the finger articulation.

The computational complexity can be reduced by observing that the hand motion is highly constrained. Previous work has shown that by incorporating motion constraints, the dimensionality of the feasible space can be significantly reduced [10, 11, 15]. To take advantage of the motion constraints for the model based approach, we must address two key issues: 1) the representation of the feasible configuration space, and 2) an efficient and effective searching algorithm associated with this representation. Previous work [3, 4, 8, 6] generally represents the manifold using piecewise linear assumption. A clustering algorithm is first applied to identify locally similar patches, which is then approximated by a linear manifold in a lower dimensional space determined by PCA or other dimensionality reduction techniques. For the case of object recognition, an affinity measure is defined between manifolds that would

IEEE COMPUTER SOCIETY

maximize the class separability[3, 8]. For tracking, a transition probability table is constructed to model the motion dynamics[4, 6]. It is generally a difficult problem in determining a suitable clustering algorithm and to construct an accurate manifold approximation.

In this paper, we propose a novel representation for the feasible configuration space using a set of discrete samples collected from CyberGlove. Each sample corresponds to one set of joint angle parameter that defines the hand shape. By using the entire set of samples directly, we do not attempt to find an approximated parametrization of the lower dimensional manifolds embedded in the feasible space, thereby avoiding the estimation error due to incorrect assumptions and lossy dimensionality reduction techniques.

For searching in this discrete space, we propose to use a stochastic Nelder-Mead (NM) simplex search. NM method is a classical direct search algorithm that is used for the cases when gradients can not be accessible or evaluated. However, like all other bottom-up approaches, it is often trapped in local minima for nonconvex objective functions. Therefore we propose to combine the top-down multiple hypothesis approach and estimate the optima with several simplices.

In Sec 2 we describe the representation of the feasible configuration space in greater details. Sec 3 gives the basic form of NM simplex search and how we adapt it to search in our new representation. Then multiple hypothesis version of NM simplex is presented in Sec 4. Sec 5 and 6 describe how we combine global and local motion and how we obtain the objective function values. Finally, experiment results are shown in Sec 7 and conclusions are given in Sec 8.

2 Non-Parametric Representation of the Hand Configuration Space

The hand motion consists of global M_G and local M_L finger motions. Global motion includes the 3D translation $\mathbf{t}$ and rotation $\mathbf{R}$ of the entire hand. M_L is represented by the set of joint angles θ_i, which has roughly 20 degrees of freedom (DOFs) [10]. Exhaustive searching in this space without any prior knowledge of the feasible space is a nearly impossible task. How to model natural configuration distribution of the feasible states is the key in successfully reducing the computational complexity. One way to model the feasible space is by observing the piecewise smooth linear manifold structures that lie in a lower dimensional space. Wu *et al.* [24] learned the local manifold structure by collecting real hand motion data and apply PCA and other anatomical constraints to reduce the dimensionality to 7. Observing the linear manifold for motion representations, they were able to effectively track finger motions. Stenger[20] on the other hand, also made use of the linear manifold approximation to help generate templates for constructing the tree structure representation of the configuration space. However, to obtain a good manifold parametrization is generally difficult. Furthermore, PCA can only model the global characteristics. It is not capable of identifying the local manifold structures and their true dimensionality.

We propose to adopt a nonparametric representation and model the entire feasible space directly from the set of N collected CyberGlove data. The entire feasible space Ψ is defined by the set of $\theta_i, i = 1 \ldots N$, 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 representation of Ψ. Therefore, this approach avoids the error induced by incorrect approximations of the representation. Furthermore, the motion constraints are automatically embedded in this discrete model. Clearly, the model can be 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

This section presents the basic Nelder-Mead search method, and the adaptation for searching in the feasible configuration space Ψ defined in sec. 2

3.1 Nelder-Mead Simplex Algorithm

With the feasible configuration space Ψ defined, a search algorithm must be designed that is appropriate for this structure. Given the estimated state x_t^* at time t, the goal is to identify x_{t+1}^* which minimizes certain objective function $f(x)$. However, 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 estimate the derivative of 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 which require first or second derivatives, such as Newton methods.

Having only the access to the function values $f(x)$, one alternative is the Nelder-Mead (NM) search method [21] which belongs to the class of direct search methods. The NM method attempts to minimize a scalar-value nonlinear function of n real variables using only function values, without any derivative information, whether explicit or implicit. 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 $\mathcal{R}^n$.

Through a sequence of elementary geometric transformations, the initial simplex moves, expands, or contracts towards the minimum. At each step, The worst vertex with highest cost $x_{max} = \arg\max_{x \in S} f(x)$ is replaced by one with smaller function value. In the case of visual hand tracking, the objective function is defined as the negative of the likelihood function $-p(\mathbf{z}|x)$, where $\mathbf{z}$ is the image observation and x is the hand state.

The procedure for performing NM search at each iteration is described in the following steps (See [21] for details). At the beginning of each iteration, the worst vertex x_{max} is selected and the centroid $\overline{x} = \frac{1}{n}(\sum_{i=0}^{n} x_i - x_{max})$ is computed. Then depending on $f(x_r)$, we perform the following operations to obtain the new vertex x_{new}which replaces x_{max}.

- **Reflect** $x_r = (1+\alpha)\overline{x} - \alpha x_{max}$.
- **Expand** $x_e = \gamma x_r + (1-\gamma)\overline{x}$.
- **Contract** $x_c = \beta x_{max} + (1-\beta)\overline{x}$

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

Given the hand state θ_t, the NM method begins by generating an initial simplex S_t^0 around θ_t, and the iterative procedure guides the search towards a minimum. Although NM method is suitable for searching when the gradient is not easily accessible, the basic form of NM does not take advantage of the motion constraints embedded in the feasible space Ψ. To incorporate the constraints in the search process, instead of performing the unconstrained simplex search in the continuous domain, we propose a two stage hierarchical NM search. In the coarse level, we 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 section 3.1, and a nearby configuration θ' is located to replace $\theta_{max} \in S_t^{k+1}$.

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

There are many algorithms and data structures designed for locating a nearest sample point in a set from any given location, such as Voronoi diagram, kd-tree, and ANN. In our experiment, we implemented the kd-tree structure. Since NM method will converge towards a minimum, the nearest neighbor does not need to be very exact. By constraining the searching to the discrete space Ψ, the hand motion constraints are automatically applied to 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, a second stage is needed to refine the search. In the first stage, the NM method begins with a larger simplex and ends with a more relaxed termination condition in the discrete domain. Then in the second stage the algorithm continues the iteration in the continuous domain with a more strict termination condition.

4 Stochastic Simplex Tracking

Like all bottom-up search algorithm, the NM simplex algorithm presented in the previous section fails when the cost function has multiple minima. Because of the noise presented in image feature extraction and the nontrivial definition of the cost function (see Sec 6), the cost function can not be convex. Furthermore, the result of the search algorithm outputs only the maximum likelihood estimate $x_t^* = \arg\max_x p(\mathbf{z}_t|x)$ when the cost function is defined as $f_t(x) = -p(\mathbf{z}_t|x)$.

One way to tackle the multiple minima problem is to employ the multiple hypotheses approach. In the tracking literature, particle filtering has been widely used as an effective multiple hypotheses tracker to reduce ambiguities. The particle filter [1, 9] uses a probabilistic approach that estimates x^* from a time evolving pdf $p(x|\mathbf{z})$ through the Bayes formulation:

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

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. The posterior density $p(x_t|\mathbf{z}_t)$ is represented using Monte Carlo simulation, with a set of n random samples and weights $\{s_t^{(n)}, \pi_t^{(n)}\}$ to approximate arbitrary nonlinear multi-modal pdf. However, particle filtering techniques becomes impractical for high dimensional state space as the number of samples required grows exponentially with respect to the number of dimensions. To cope with this problem, Cham[5] proposes to use a semi-parametric representation, and Wu[24] uses importance sampling from an underlying distribution. Another problem with the particle filter is the degeneracy phenomenon where sample weights become insignificant over time [1].

We propose a stochastic simplex search algorithm by combining the NM algorithm with the particle filtering tracking scheme. 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 at t to locate the new modes at $t+1$ through NM method. The new algorithm combines the advantages of each approach to reduce the limitations induced by employing NM or particle filtering 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 included in estimating a more accurate current hand state. Third, since each simplex will converge to a local minimum, the converged samples will have more significant weights. Cham [5] also employed an iterative Gauss-Newton method to locate the modes of the likelihood $p(\mathbf{z}_t|x_t)$. In our case, since we do not have access to the gradient, NM simplex search is employed.

To approximate the tracking prior $p(x_{t+1}|\underline{\mathbf{z}}_t)$, we first draw random samples $\tilde{x}_t^i$ from $p(x_t|\underline{\mathbf{z}}_t)$. Then an initial simplex S^{i0} is generated from each $\tilde{x}_t^i$, which corresponds to a mode in $p(x_t|\underline{\mathbf{z}}_t)$. Next the two stage simplex search is carried out to obtain a local minimum corresponding to a mode of the pdf:

$$S^{i*'} = \mathcal{NM}^d(S^{i0})$$
$$S^{i*} = \mathcal{NM}^c(S^{i*'}) \qquad (2)$$

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

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

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:

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

$\pi_{t+1}^i = p(\mathbf{z}_{t+1}|x_{t+1})$ is computed as described in sec. 6. This procedure is similar to the result of $p(x_{t+1}|\underline{\mathbf{z}}_t) = \int_{x_t} p(x_{t+1}|x_t)p(x_t|\underline{\mathbf{z}}_t)dx_t$ where the dynamics $p(x_{t+1}|x_t)$ is often modelled as a Gaussian. In the case of NM, the iterative procedures (Sec 3.1) automatically drives the simplex towards a new mode.

5 Global and Local Motions

In [22], Wu proposed to simultaneously estimate both global and local hand motions by a divide-and-conquer approach. Rather than estimating the optimal state parameters altogether, the global motion and local motion parameters are estimated separately and combined in an iterative manner. Different algorithms can be used to independently estimate each set of parameters, such as the GA-based algorithm used in [22] and the iterative closed point algorithm employed in [12]. In our experiments, we apply the stochastic NM simplex search in the continuous space for global parameter estimation. Then the two stage simplex search (Sec. 4) is employed to recover the finger motion. These two steps are then repeated until the results converge. Our experiments have shown that this approach reduces the computational complexity while effectively estimates the hand motion.

6 Model Matching

We employ edge and silhouette observations to measure the likelihood of hypothesis as in [12]. The cylinder hand model is first projected onto the image plane as described in [19] to obtain the edge points that define the projected shape. Assume K projected model samples are generated, edge detection is performed on the points along the normal of these samples. Assuming that M edge points $\{z_m, m = 1, \ldots, M\}$ are observed, and the clutter is a Poisson process with density λ, then,

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

We notice 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 projected 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, we define the objective function at time t to be the negative of the likelihood function:

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

7 Experiments

In our experiments, we use a 3D hand model and each finger phalanx is represented using a truncated cylinder. Hand projection is generated as described in [19] and is superimposed on the real hand image. All experiments are performed in a cluttered background. The first experiment demonstrates the robustness of our 3D model-based tracking in which the hand undergoes both translation and rotations (Figure 1). The sequence shows that the 3D model can handle significant occlusions. The fingers are assumed rigid in this case. We use 10 simplex search for each frame.

In the second video sequence, the fingers bend and extend while the hand moves simultaneously (Figure 2). This experiment shows the robustness of our articulation tracking algorithm. We use 30 simplices for finger articulation tracking and 10 simplices for global motion. We have also tested the sequence using CONDENSATION algorithm with 5000 samples, and the algorithm fails after about 10 frames.

In addition to the superimposed model projection, 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.

8 Conclusions

This paper proposes to track the articulate hand motion by addressing two key issues: 1) the representation of the feasible configuration space and 2) an efficient tracking algorithm associated with the representation. We propose to directly model the feasible space from real hand motion data. The advantage of this representation is that it automatically incorporates the motion constraints in this model. Furthermore, the errors induced from the approximation algorithms can be avoided.

In order to utilize this discrete representation for tracking hand motion, we propose a stochastic NM simplex search algorithm which is modified to work for the discrete space. One main advantage of the NM method is that it does not require the knowledge of gradients, which in our case is difficult to obtain. Since direct search methods are often trapped in local optima, we incorporated particle filtering framework with the simplex search. The experiment results show that our algorithm is robust in tracking hand motions in cluttered background.

Acknowledgments

The authors thank Howard Huang for his numerous insightful comments. 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] V. Athitsos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume II, pages 432–439, Madison, 2003.

[3] 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.

[4] M. Brand. Shadow puppetry. In *Proc. of IEEE Int'l Conf. Computer Vision*, pages 1237–1244, 1999.

[5] T.-J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 239–245, 1999.

[6] 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.

[7] 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.

[8] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman. Clustering appearancs of objects under varying illumination conditions. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume I, pages 11–18, 2003.

[9] 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.

[10] 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.

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

[12] J. Lin, Y. Wu, and T. S. Huang. Capturing human hand motion in image sequences. In *Proc. of IEEE WMVC*, pages 99–104, 2002.

[13] 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.

[14] 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.

[15] 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.

[16] R. Rosales, S. Sclaroff, and V. Athitsos. 3D hand pose reconstruction using specialized mappings. In *Proc. IEEE Int'l Conf. on Computer Vision*, Vancouver, Canada, July 2001.

[17] J. Segen and S. 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.

[18] 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.

[19] 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.

[20] B. Stenger, A. Thayananthan, P. Torr, and R. Cippola. Filtering using a tree-based estimator. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

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

[22] 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.

[23] Y. Wu and T. S. Huang. View-independent recognition of hand postures. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume II, pages 88–94, 2000.

[24] 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.

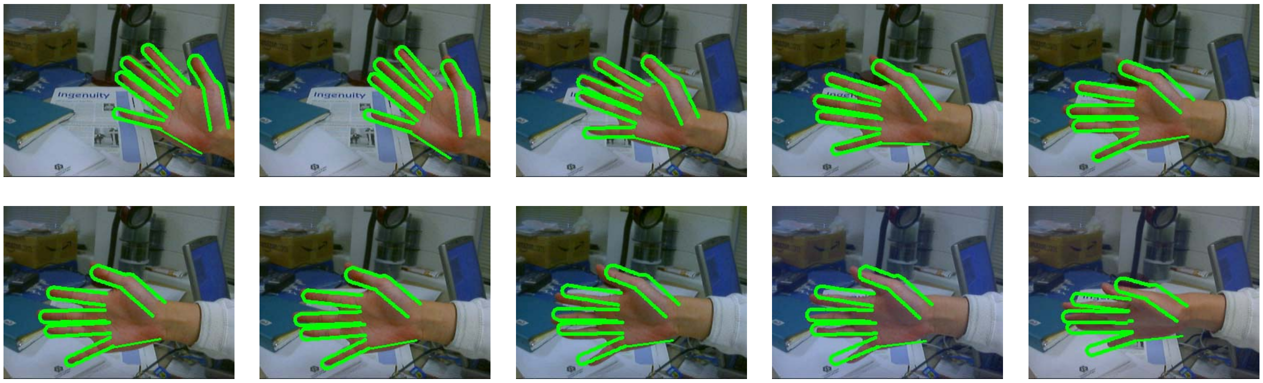

Figure 1: Tracking global hand motions involving translation and out-of-plane rotation. The projected model edge points are superimposed on the real hand image.

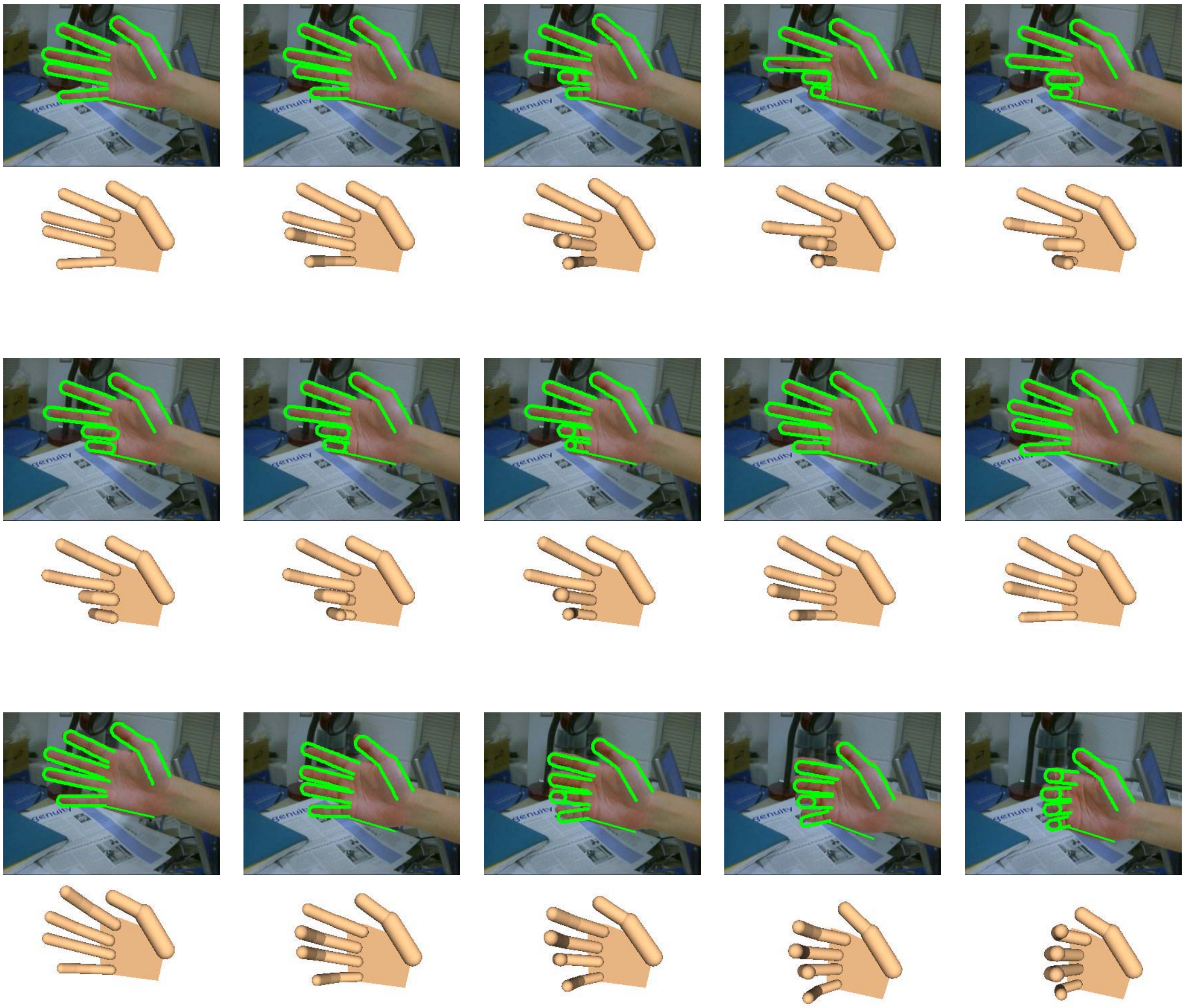

Figure 2: Simultaneously tracking finger articulation and global hand motion. The projected edge points are superimposed on the real hand image. Below each real hand image, a corresponding reconstructed 3D hand model is shown for better visualization.