3D Hand Reconstruction from a Monocular View

Sung Uk Lee, Isaac Cohen
Institute for Robotics and Intelligent Systems
Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-0273
{slee|icohen}@usc.edu

Abstract

In this paper, we describe an approach for human hand motion detection and its reconstruction using an articulated model with hand kinematics constraints. Each finger can be modeled as planar robot arm with 3 joints and 3 links. We assume that first joints connecting palm and each finger configure a rigid form. By this simplification, we have several hand constraints which reduce the complexity of the estimation process and allow to infer the 3D hand motion and pose in ambiguous situations. The main advantage of the proposed approach is its ability to capture general articulated hand motion with self self-occlusion and rotation of the palm. The proposed method is illustrated on a set of examples of a hand motion captured from a monocular image sequence.

1. Introduction

The need for an accurate hand motion tracking has been increasing in HCI and human sensing related applications. Hand gestures can be used as a convenient user interface for various computer applications in which the state and the action of the users are automatically inferred from a set of video cameras. HCI related applications remain the most popular however; we have focused our effort on the accurate detection and tracking of un-instrumented hands for assessing user performance in accomplishing a specific task. Assessing hand-eye coordination for post-trauma rehabilitation has emerged as one that displays the most challenging issues such as: accuracy, complexity of the gesture and a large amount of self-occlusions.

1.1. Related Work

A large number of studies have been proposed for capturing human hand motion. Some of them have focused on very specific domains and thus have a large number of restrictive assumptions. In [9][10][11], trajectories of hand motion are considered for classifying human gestures and are sufficient for the studied domain. Several frameworks have been proposed we can distinguish two types: appearance-based and model-based methods. In [7] an appearance-based method is proposed for characterizing the various hand configuration and finger pose. This relies on a learning-based approach and requires training the algorithm on a large data set that depicts all possible configurations. Similarly in [8], articulated pose of a hand is formulated as a database indexing problem. From a large set of synthetic hand images, the closest matches of an input hand image are retrieved as an estimated pose.

Model based approaches rely on the physical properties of the hand and finger motion [1][2][3] in order to infer the hand chararcteristics (i.e. joints articulations) from a single or multiple views. Huang et al. [5] proposed a model-based approach that integrates constraints of natural hand motion. They also proposed a learning approach to model the kinematics constraints directly from a real input device such as the CyberGlove [4]. They focused on the analysis of local finger motion and constraints from observed data. For articulated motion tracking, a Sequential Monte Carlo algorithm based on importance sampling is employed. In a fixed viewpoint (i.e. the palm is not moving), this produces good results; however, global hand position and motion are not dealt with.

Recently, some studies of integrating various cues such as motion, shading and edges were proposed. Metaxas et al [6] compute the model driving forces using gradient-based optical flow constraint and edges information. A forward recursive dynamic model is introduced to track the motion in response to 3D data derived forces applied to the model. They assume that global hand motion is a rigid motion. No simultaneous articulated motion is expected. When the fingers are in motion of significant relative motion with occlusions, they hardly derive the hand model.

1.2. Overview of Proposed Approach

We present in this paper a hand pose estimation technique based on the matching of an articulated hand model to the detected hand silhouette. Inferring hand motion from monocular view is challenging as the detected 2D silhouette is not sufficient to depict a complicated hand pose. Several studies have addressed this problem by using an articulated hand model. Although these methods can track the articulated hand motion from monocular view, many of them are computationally too expensive and have difficulties in addressing hand's self-occlusions. The proposed articulated hand model efficiently tracks detected hand in the presence of self-occlusions and has a low computational cost. Starting with a segmentation of the hand silhouette from the video stream, an articulated hand model is fitted onto the 2D mask of the hand. For initialization, the 3D hand model configuration defined by the joint offset and size are estimated refined by the favorable stretched-out hand pose. Section 3.1 describes the initialization process in detail. First, the joints connecting palm and each finger form a rigid shape. The corresponding 3D model is fitted and refined. A global hand pose is computed using this rigid form constraint and then estimation of each finger motion is performed. To

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

reconstruct accurate pose, 3D model is updated by re-projecting it on 2D detected silhouette in the tracking process. An overview of the proposed approach is shown in figure 1.

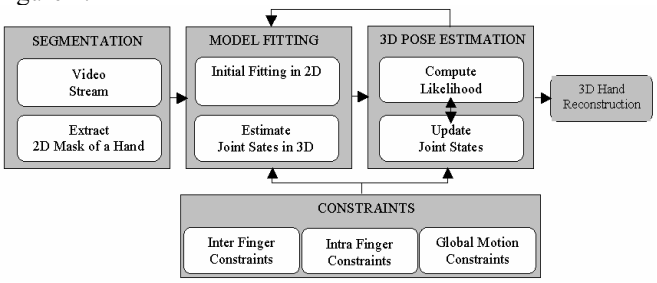

Figure 1. *Overview of the proposed approach.*

2. 3D Hand Model and Constraints

A human hand consists of many joints and links, and it can generate numerous motions. Although the pose of a hand seems infinite, general hand motion is highly constrained. In a finger, each joint has a relationship with other joints. The configuration of a finger is also constrained by neighboring finger configurations. By modeling human hand and its constraints, we can reduce the computational complexity of search/matching problem and we can even estimate the 3D hand configuration from a single 2D image.

2.1. Articulated Hand Model

The articulated hand model we use is shown in figure 2. It consists of 15 joints and has 20 degrees of freedom (DOF). Each finger can be modeled as a typical planar robot arm [12] consisting of 3 joints and 3 links.

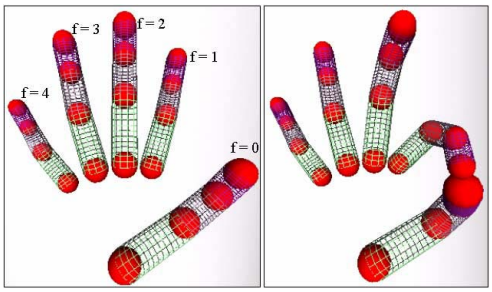

Figure 2. *The articulated hand model used in this work consists of 15 joints and 20 DOF.*

Each joint j_i^f is a revolute joint where f is the index of fingers enumerated from thumb to little finger ($f = 0 \sim 4$). All the first joints (where $i = 0$) connecting palm and each finger have *2* degrees of freedom and other joints have a single rotational axis. Therefore, the suggested model has 20 DOF. Each link can be represented as a cylinder connecting each joint with variable radius. The initial thickness and length of the cylinders and joints are defined from an average male hand, and they are automatically estimated in the initialization step.

2.2. Modeling Hand Constraints

Without the use of any physical constraint inherent to the hand, modeling and analyzing, the hand motion can be quite complex. However, joints in a hand are interrelated and their motions are related by physical constraints. Huang et al [4] showed in detail that human hand motion is highly constrained. A learning approach was proposed to model the hand configuration and characterize the physical constraints space directly sensed from a CyberGlove. In this paper, three types of constraints are defined for natural hand motion.

Intra-Finger Constraints. In this category, the main constraint considered is of finger joints during a bending motion. The number of rotational axes in a joint is limited. By using the planar 3-joint-arm finger model, the DOF of a finger are reduced to 4. The first joint of a finger j_0^f has 2 rotation axes – one for bending motion and another is for abduction /adduction motion – but remaining joints have single rotation axis. Table 1 shows the range of each joint angle for the bending motion.

Table 1. *Joint angle constraints for bending motion*

j_i^f	$i = 0$	$i = 1$	$i = 2$
$f = 0$	$-10^o \sim +45^o$	$+0^o \sim +90^o$	$-5^o \sim +90^o$
$f = 1 \sim 4$	$-15^o \sim +90^o$	$-5^o \sim +90^o$	$-5^o \sim +90^o$

Each contiguous joint has the following constraint:

$$\angle j_i^f = \alpha_i \left| \angle j_{i-1}^f \right| \qquad (1)$$

where finger index $f = 1$ to 4 and joint index $i = 1, 2$. α_i = (limit angle of j_i^f) / (limit angle of j_{i-1}^f). This type of constraint is not strictly applied, but is an efficient guide of joint offset estimation.

Inter-Finger Constraints. Proposed model has fixed geometric proportion among j_0^f (offsets of all first joint of a finger). No joint can overlap each other in 3D coordinates. Interrelated motion constraints for abduction/adduction motion between adjacent fingers are defined in this category. Adjacent finger tends to follow the dominant articulated motion.

Global Motion Constraints. This category defines some restrictions on hand motion. The objective is tracking natural hand motion, where the starting hand pose is a favorable pose for initialization. No external torque on a joint is expected and hand motion should be smooth. Within these constraints, we address the problem of tracking articulated hand motion with self-occlusion, rotation and translation. The proposed approach does not address the scaling and depth estimation from monocular view. Some recent studies introduce similar constraints to this paper. However, they are hardly explicitly modeled by equations.

3. 3D Hand Model Fitting with Constraints

3.1. Initial Model Fitting in 2D

A sequence of hand motion begins with a favorable initial pose as shown in figure 3. By using an ellipsoid fitting on top of the segmented hand silhouette, an estimated scale and orientation of a hand are computed. Borgefors' two-pass algorithm is used to compute the distance map *d(x,y)* of a hand image for refining the model position to the observed silhouette. The distance map is

defined as follows:

$$D(x,y) = \begin{cases} -d(x,y), & (x,y) \in Silhouette \\ d(x,y), & Otherwise \end{cases} \quad (2)$$

By minimizing the distance $D(x,y)$ along the gradient $\nabla D(x,y)$, each joint offset can be refined. Each joint width (thickness) can be also estimated with $D(x,y)$. Figure 3 shows an example of model initialization in 2D. Based on this information, the initial configuration of 3D hand model is refined.

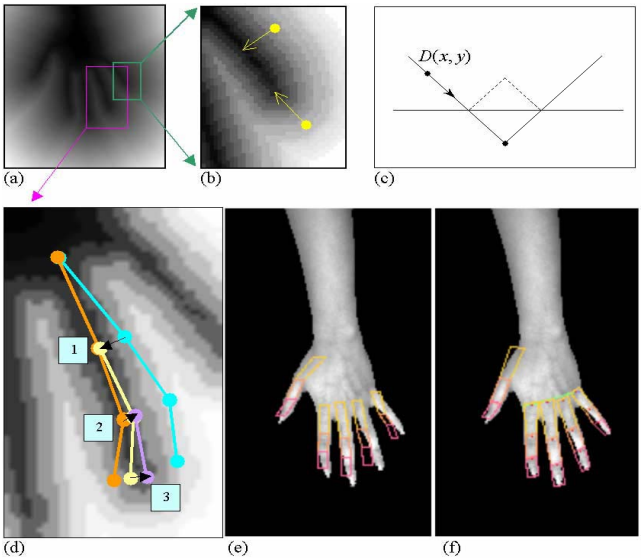

Figure 3. *Model initialization. (a) Distance map. (b), (c) Approaching median axis (minimum distance) along the gradient of distance map. (d) Joint position refinement. (e) Coarse 2D model with ellipsoid fitting. (f) An example after refinement.*

3.2. 3D Pose Estimation from a 2D Image

Using inter-finger constraints and global motion constraints, a hand motion can be divided into global pose and individual finger motion.

3.2.1. Global Pose Estimation

At time *t*, first joints of j_0^0, j_0^1, and j_0^4 form a palm-triangle *PT(t)* which has three sides of w_t^0, w_t^1, w_t^2. In the two consecutive palm-triangle *PT(t)* and *PT(t+1),* the proportion of corresponding sides of each palm-triangle provides the global pose. By the global motion constraint, j_0^0 is the origin of a hand configuration. Thus, the translation vector T_g can be estimated by two consecutive 2D positions of j_0^0 detected from the hand silhouette. The rigid motion constraint, w_t^p / w_{t+1}^p gives the angle for rotation vector R_g. In figure 5, palm-triangle of bright violet shows the estimation results of global pose.

3.2.2. Individual Finger Motion Estimation

After estimating the global rotation and translation [R_g, T_g], each joint angle of a finger can be estimated from the equations given in figure 4. The length of each link of a finger is estimated at the initialization and remains constant. This allows inferring the joint angle θ_i for the bending motion.

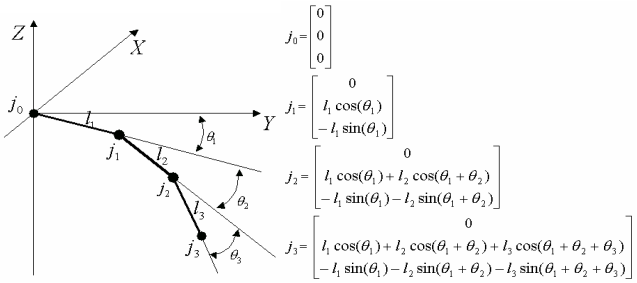

Figure 4. *Kinematics of the finger model.*

4. Articulated Motion Tracking

4.1. Likelihood Computation

The likelihood measurement for fitting and tracking the articulated model is based on two types of features, the finger boundary and finger region. Given a predicted pose, synthesized appearance of hand model is re-projected onto 2D image. Likelihood measure is computed by matching the boundary of the model with the extracted foreground. The likelihood for boundary matching is expressed as:

$$\boldsymbol{L}_b = \prod_{j=1}^{Nc} P(C_j / S) \quad (3)$$

The similarity measure is given by,

$$P(C_j / S_k) = \exp[-d^2(C_j, S_k) / \sigma^2], \quad (4)$$

where σ^2 is the disparity variance. For each hand configuration, the position of each joint segment is defined by $S_t^i = \{S_1, ..., S_{Ns}\}$, where the number of joint segment *Ns* is 3x5 =15. Each contour point C_j is matched to the closest segment S_k; and $k = \arg\min_l d(C_j, S_l)$ where $d(C_j, S_l)$ is the distance from C_j to the edge of segment S_l. Likelihood measurement for region matching is: $\boldsymbol{L}_r = (P_\alpha)^{Na}$, where *Na* is the number of pixels lie outside the valid finger region. P_α is the false negative error in foreground segmentation. Distance map $D(x,y)$ and the gradient $\nabla D(x,y)$ are efficient to check the validity of a pixel. The combined likelihood for matching the articulated hand model to the detected silhouette is given by: $\boldsymbol{L} = \boldsymbol{L}_b \times \boldsymbol{L}_r$, More detail description can be found in [13].

4.2. Motion Tracking

The proposed tracking technique is based on '*tracking by synthesis*'. To maximize the likelihood, possible sets of configuration is tested. The adjustment of a hand model is guided by the motion constraints and the configuration of the hand at previous frame. Disparity vector of the closest boundary between the model segment and the detected

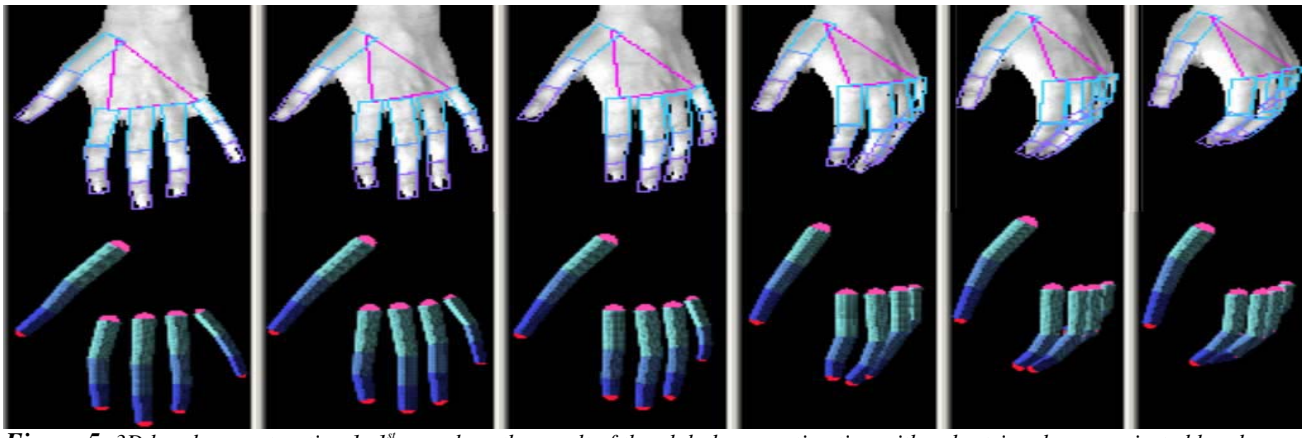

Figure 5. *3D hand reconstruction 1: 1st row show the result of the global pose estimation with palm-triangle. re-projected hand model 2nd row shows the reconstructed 3D hand model.*

silhouette is computed for estimating the position of a joint angle. The adjusted 3D hand configuration is projected back onto the 2D image and updated from 2D image measurements back and forth. Figure 5 shows some results of motion tracking. Reconstructed 3D hand model and its re-projected 2D model are correctly matched with the detected silhouette. Various hand poses including self-occlusion and the corresponding results are shown in figure 6.

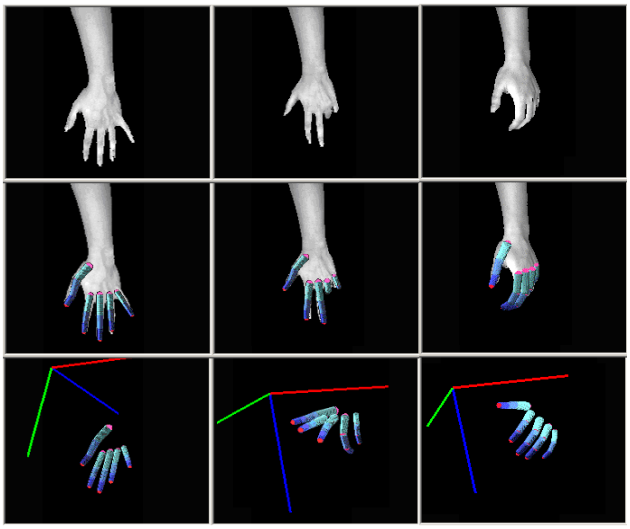

Figure 6. *3D hand reconstruction 2: 1st Row shows the original hand pose. 2nd row represents the reconstructed pose. 3rd row is different angle view.*

5. Conclusion

In this paper, an articulated hand model and its constraints were introduced. We have mainly addressed the problem of tracking the simultaneous motion of rotation, translation, and self-occlusion in monocular view. Proposed approach is easy to extend for real-time and multi-view framework, although the target motion is somewhat restrictive. Our next work will focus the development of an efficient joint angle tracking technique with large occlusions due to objects manipulated by the user.

Acknowledgments

This research was partially funded by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, under Cooperative Agreement No. EEC-9529152.

6. References

[1] J.M. Rehg and T. Kanade, "Model-Based Tracking of Self-Occluding Articulated Objects", *ICCV95*, pp.612-617, 1995.

[2] Q. Delamarre and O. Faugeras, "Finding pose of hand in video images: a stereo-based approach", *In IEEE Conference on Automatic Face and Gesture Recognition*, Japan, 1998

[3] B. Stenger, P. R. S. Mendonça, and R. Cipolla, "Model-Based 3D Tracking of an Articulated Hand", *In CVPR*, Volume II, pages 310-315, Kauai, USA, December 2001.

[4] John Lin, Ying Wu and Thomas S. Huang, "Modeling Human Hand Constraints", *In. Workshop on Human Motion* , Austin, Dec., 2000.

[5] Ying Wu, John Lin and Thomas S. Huang, "Capturing Natural Hand Articulation", *in ICCV'01*, Vol.II, pp.426-432.

[6] S. Lu, G. Huang, D. Samaras and D. Metaxas, "Model-based Integration of Visual Cues for Hand Tracking" In Proc. *WMVC 2002*

[7] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff, "3D Hand Pose Reconstruction Using Specialized Mappings" *IEEE ICCV*. Canada. Jul. 2001.

[8] V. Athitsos and S. Sclaroff, "Estimating 3D Hand Pose from a Cluttered Image", *In CVPR 2003*, USA, June 2003.

[9] R. Cutler and M. Turk, "View-based interpretation of realtime optical flow for gesture recognition", *In Proc. Face and Gesture Recognition*, pp. 416–421, 1998.

[10] A. Nishikawa, A. Ohnishi, and F. Miyazaki. "Description and recognition of human gestures based on the transition of curvature from motion images", *In Proc. Face and Gesture Recognition*, pp. 552–557, 1998.

[11] Y. Xiong and F. Quek, "Gestural Hand Motion Oscillation and Symmetries for Multimodal Discourse: Detection and Analysis", *Workshop on Computer Vision and Pattern Recognition for HCI.* June, 2003.

[12] L. Sciavicco, B. Siciliano, "Modeling and control of robot manipulators" MacGraw-Hill, New York, 1996.

[13] Mun Wai Lee, Isaac Cohen, Soon Ki Jung, "Particle Filter with Analytical Inference for Human Body Tracking", *IEEE Workshop on Motion and Video Computing*, 2002.