Human Hand Modeling, Analysis and Animation in the Context of HCI

Ying Wu, Thomas S. Huang
Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{yingwu, huang}@ifp.uiuc.edu

Abstract

The use of human hand as a natural interface device serves as a motivating force for research in visual analysis of highly articulated hand movement. Since hand motion covers a huge domain, the scope of this paper is limited to the developments of 3D model-based approaches. Numerous 3D models that have been used to analyze hand motion are studied. Various approaches to articulated motion analysis are discussed. Some realistic synthesis methods are also included in this paper. We conclude with some thoughts about future research directions.

1 Introduction

Through the evolution of user interfaces (UI), keyboards were the primary devices in text-based UI and mice moved us to GUI. What will be the counterpart of the mouse when we are trying to explore 3D virtual environments (VE) in Human Computer Interaction (HCI) or Perceptual User Interfaces (PUI)? At least, 3D input should be supplied. In current VE applications, keyboards, mice, wands and joysticks are still the most fundamental, popular and dominant controlling and navigating devices. However, they are inconvenient and unnatural. In many cases such as CAVE-like VEs, magnetic trackers are being used as sensors of 3D inputs with some of these devices. However, they are prone to magnetic interference, and unable to give the feeling of immersiveness.

The use of hand gestures has become an important part of HCI in recent years [1, 8, 12]. In order to use human hands as a natural interface, some alternatives, such as glove-based devices, are used to capture human hand motion by attaching some sensors to measure the joint angles and spatial positions of hands directly. Unfortunately, such devices are expensive and cumbersome.

Non-contact vision-based technique is one of the promising alternatives to capture human hand motion by affordable camera settings, which serves as a motivating force for research in the modeling, analyzing, animation and recognition of hand gestures.

According to different application scenarios, hand gestures can be classified into several categories such as conversational gestures, controlling gestures, manipulative gestures and communicative gestures. Sign language is an important case of communicative gestures. Since sign languages are highly structural, they are very suitable for acting as a test-bed for vision algorithms [15, 18]. At the same time, they can also be a good way to help the disabled. Controlling gestures are the focus of current research in vision-based interface (VBI) [16]. Virtual objects can be located by analyzing pointing gestures [12]. Some display-control applications demonstrate the potential of pointing gestures in HCI [2]. Another controlling gesture is the navigating gesture. Instead of using wands, the orientation of hands can be captured as an 3D directional input to navigate the VEs [11]. The manipulative gesture will serve as a natural way to interact with virtual objects [7]. Tele-operation and virtual assembly are good examples of applications. Communicative gestures are subtle in human interaction, which involves a lot of psychological studies, however, vision-based motion capturing techniques can help those studies [21].

In this paper, we study 3D hand models employed in current research. Various articulated motion analysis approaches are discussed. Some realistic synthesis methods are also included in this paper. We conclude with some thoughts about future research directions.

2 Hand Modeling

The human hand consists of many connected parts forming kinematical chains so that hand motion is highly articulated. At the same time, there are many constraints among fingers and joints that make the dynamics of hand motion even harder to model. Usu-

0-7803-5467-2/99/ $10.00 © 1999 IEEE

ally, hand can be modeled in several aspects such as shape, kinematical structure, dynamics and semantics. Hand models are not only used in hand animation applications, but also employed to analyze hand motion using the approach of *"analysis-by-synthesis"*. Different models are suitable for different HCI applications. In animation application, the shape model should be as fine as possible and the motion model should be as realistic as possible, but in motion analysis, a simple kinematical model is often adequate. However, if the hand models of different aspects can be integrated, hand motion analysis may hopefully be investigated in a comprehensive way.

2.1 Modeling the Shape

Hand shape models can be classified into several groups such as geometrical models, physical models and statistical models.

Spline-based geometrical surface models represent a surface with splines to approximate arbitrarily complicated geometrical surfaces. These spline-based surface models can be made as realistic as possible, but many parameters and control points need to be specified [9]. An alternative is to approximate the homogeneous body parts by simpler parameterized geometric shapes such as *generalized cylinders* or *super-quadrics.* The advantage of this method is that it can achieve equally good surface approximation with less complexity [14, 3]. Other than parametric models, free-form hand models are defined on a set of 3D points [6]. Polygon meshes that are formed by those 3D points approximate the hand shape, which is computationally efficient. Cyber Scanner, MRI techniques or other space digitizer may be used to get the range data directly [6]. Another way is to reconstruct the hand model from multiple images of different views.

Physical hand shape models emphasis the deformation of the hand shape under the action of various forces [18]. The motion of the model is governed by Newtonian dynamics. The internal forces are applied to hold the shape of the model, and the external forces are used to fit the model to the image data. Examples are simplex mesh model [6] and finite element method model [17].

Statistical hand shape models learn the deformation of hand shape through a set of training examples that can be 2D images or range images. Mean shape and modes of variation are found using PCA. A hand shape is generated by adding a linear combination of some significant modes of variation to the mean shape. Point distribution model is a good example [6].

There are some related issues such as model alignment and pose estimation. Model alignment is to establish correspondences between the image and the 3D model. Pose estimation is a special case of model alignment. Given image points and corresponding model points and camera intrinsic parameters, pose determination is to find a rotation $\mathbf{R}$ and translation $\mathbf{t}$ to align the model and image by:

$$minimize \sum_{i=1}^{N} ||\mathbf{X}_i - \mathbf{P}(\mathbf{R}\mathbf{x}_i + \mathbf{t})||^2 \qquad (1)$$

where $\mathbf{P}$ is the projection transformation which is given by the intrinsic parameters of the camera, $\mathbf{X}_i$ is image point and $\mathbf{x}_i$ is model point.

2.2 Modeling the Kinematical Structure

If the hand is treated as a set of sub-objects, not only should each sub-object be modeled separately, the kinematical relations among the sub-objects should also be included in the hand model. The skeleton of a hand can be abstracted as a stick figure so that the dimension of each sub-object is reduced to its *link length.* The name for each joint is indicated in Fig.1.

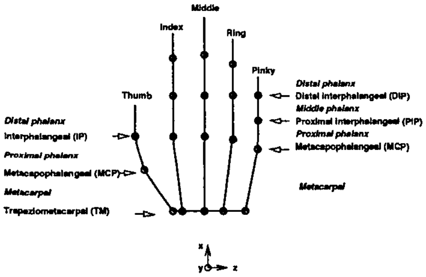

Figure 1: Hand skeleton structure

Each finger is modeled as a kinematical chain with the palm as its base reference frame. The fingertip is the end-effecter of the kinematical chain that is formulated as:

$$\mathbf{x}^b = \mathbf{H}_0^b(\ \theta_{MCP\_AA})\mathbf{H}_1^0(\ \theta_{MCP})\mathbf{H}_2^1(\ \theta_{PIP})\mathbf{H}_3^2(\ \theta_{DIP})\mathbf{x}^3 \qquad (2)$$

where $\mathbf{x}^3$ is the fingertip in DIP frame, while $\mathbf{x}^b$ is fingertip in the base frame. $\mathbf{H}_i^j$ is the coordinate transformation which transform the i frame to the j frame.

Given the coordinates of the fingertip in its local frame and reference frame, the *inverse kinematics problem* is to find the joint angles. Generally, gradient-based methods are taken to solve this problem by deriving the *kinematical Jacobian* [13]. There are other alternatives in the literature such as genetic algorithm [20] and neural network methods [2]. Since it is an ill-posed problem, unique solution can only be found by adding some constraints. Joint angles should be in certain range such as $0 \leq \theta_{MCP} \leq 90^0$ and $-15^0 \leq \theta_{MCP\_AA} \leq 15^0$ [10]. At the same time, DIP and PIP joints are not independent and it can be described as $\theta_{DIP} = \frac{2}{3}\ \theta_{PIP}$ [10]. These heuristics are very important to model natural hand motion and reduce the searching space. However, constraints may be hard to find and express due to the complexity of finger motion.

2.3 Modeling the Dynamics

To capture complex hand motion and recognize continuous hand gestures, the dynamics and semantics of hand motion should also be modeled.

Kalman filtering and extended Kalman filtering (EKF) techniques are widely adopted to model the dynamics [14]. EKF works well for some tracking tasks. However, it is based on small motion assumption that often fails to hold in hand motion.

Simple hand gestures can be modeled by finite state machine [3], but it is insufficient to handle complex gestures. Considering the similarity between sign languages and spoken languages, Hidden Markov Model (HMM) and its variants are also used to model the dynamics of hand movement [15, 19]. Rule-based approach can also be applied to model the semantics of hand movements [21]. Bayesian net is another promising approach. Neural networks are also usable. These methods are essentially learning methods which learn the intrinsic dynamics from a set of training data. The knowledge of dynamics and semantics is not explicitly expressed in these methods, but implicitly stored in the structures of the learning models.

The learning results of these methods depend on the training data set, structures of learning models and training methods. One of the common problems of the learning approaches is that generalization of the learning results largely depends on the training data. However, obtaining the training samples is obviously not a trivial problem. Currently, learning dynamics (behaviors, semantics) of human motion has drawn much attention from researchers in HCI, computer vision, computer graphics, psychology etc.

3 Capturing Hand Motion

Hand motion capturing is to find the global and local motion of hand movements so that the hand posture can be recovered. Several different model-based approaches are discussed in this section.

3.1 Formulating Hand Motion

Highly articulated human hand motion consists of the global hand motion and local fingers motion, which can be expressed as

$$\mathbf{M} = \mathbf{M}(\mathbf{M}_G, \mathbf{M}_L) \tag{3}$$

where $\mathbf{M}$ is the hand motion, $\mathbf{M}_G$ is the global motion and $\mathbf{M}_L$ is the local motion. Global hand motion that presents large rotation and translation can be written as $\mathbf{M}_G = \mathbf{M}_G(\mathbf{R}, \mathbf{t})$, where $\mathbf{R}$ and $\mathbf{t}$ are rotation and translation respectively. One important issue is how to reliably track the global motion in image sequences.

Local hand motion is articulated with so many self-occlusions that make the detection and tracking hard. Local hand motion can be parameterized with the set of joint angles (or *state of hand*), $\mathbf{M}_L = \mathbf{M}_L(\ \theta)$ where θ is the joint angle set. Consequentially, hand motion can be expressed as:

$$\mathbf{M} = \mathbf{M}(\mathbf{R}, \mathbf{t},\ \theta) \tag{4}$$

One possible way to analyze hand motion is the appearance-based approach which emphasis the analysis of hand shapes in images [12]. However, local hand motion is very hard to estimate by this means. Another possible way is the model-based approach [3, 6, 9, 10, 13, 14, 18, 20]. With single calibrated camera, local hand motion parameters can be estimated by fitting the 3D model to the observation images. Multiple camera settings are helpful to deal with occlusion [10, 13, 18]. The use of a 3D model can largely alleviate the problem of *depth ambiguity* since the structure of hand is included in the model.

3.2 Selecting Image Features

In order to estimate the parameters of the model, some images features should be extracted and tracked to serve as the observation of the estimators. Hand image features can be geometric features such as points, lines, contours and silhouettes [9]. Fingertip is one of the frequently used features, because the positions of fingertips are almost sufficient to recognize some gestures due to the highly constraint hand motion [10]. Color markers are often used to help tracking the 3D position of fingertips [10, 3]. Some researchers estimate the position and orientation of fingertips by fitting a 3D cylinder to the images [3]. Line fitting is also a frequently used technique to detect the fingertips [2, 13].

Many non-geometric features are widely used as well such as color and motion. Hand can be treated as a color blob or motion blob in localization [11, 15]. There are also many studies on how to integrate multiple cues.

3.3 Capturing Hand Motion

The model-based approach in motion capturing basically is to align a model to images or even rang data by estimating the parameters of the model. This problem is closely related to the camera calibration problem and essentially a data-fitting problem that was discussed before. However, hand motion is highly articulated so that it is very hard to analyze and capture.

Different methods have been taken to analyze human hand motion. One possible way is appearance-based approaches, in which 2D deformable templates are used to track a moving hand in 2D. However, this method is insufficient to analyze and recognize hand gestures. Another possible way is 3D model-based approach, which takes the advantages of *a prior* knowledge built in the 3D models. Model-based methods track hands in 3D and recover the joint angles of hand [3, 6, 10, 9, 13, 14, 20].

One method of model-based approaches is to use gradient-based constrained nonlinear programming techniques to estimate the global and local hand motion simultaneously. Hand can be modeled as an articulated stick figure [13]. The drawback of this approach is that the optimization is often trapped in local minima. Another idea is to model the surface of the hand [3, 9, 14], and then hand configuration can be estimated using the "analysis-by-synthesis" approach. Candidate 3D models are projected to the image plane and the best match is found with respect to some similarity measurement. Essentially, it is a searching problem in very high dimensional space that makes this method computational intensive. If the surface model is very fine, an accurate estimation can be obtained. However, those hand models are user-dependent. Rough models can only give approximate estimation [14].

A decomposition method is also adopted to analyze articulated hand motion by decoupling hand motion to its global motion and local finger motion. Global motion is parameterized as the pose of the palm, and local motion is parameterized as the set of joint angles. A two-step iterative algorithm is used to find an accurate estimation. Given an initial estimation, hand pose is estimated using least median of squares (LMS) with joint angles fixed. Then the joint angles are recovered by a genetic algorithm with the global hand pose fixed. Those two steps are alternately iterated until the solution converges [20].

4 Realistic Animation and Motion Editing

Hand model can be easily animated by keyframe-based methods. If the model is driven by the set of joint angles which represents the state of hand, hand states can be interpolated by pre-specified key-frame states. If the model is driven by the position of fingertips, an inverse kinematics problem must be solved. Although it is simple to implement, the drawback of this approach is apparent. In order to obtain a realistic effect, a large number of control points must be specified along the fitting curves. To reduce the amount of motion specification, some knowledge about hand motion should be built in the animation system to execute certain aspect of movement autonomously. Some high-level control schemes and physical rules can be used to achieve this goal, however, the disadvantage is lack of interactivity.

Another realistic hand motion synthesis technique is editing captured motion [5]. The purposes of motion editing are to smooth the jerky captured motion data due to sensor noise, correct the violation of body constraints, and generate realistic movements. By this means, the animation system can alter the geometry of the hand, warp the timing, and perform seamless transition so that captured motion data can be reused. New motion can also be generated from existing motion data by motion blending.

5 Future Research Directions

In order to use human hand as a natural device in HCI, hand gestures should be tracked and understood. There are several issues related to hand modeling that need to be adequately addressed in the future. One of the aspects involves modeling the constraints between joints and fingers. Those constraints will significantly reduce the search space and lead to realistic animation. Another issue is modeling the *coarticulation* in gestures. Most current gesture applications only look into symbolic gesture commands. However, it is still hard to extract or segment those gesture commands in continuous hand movements.

Although capturing fully hand motion is not necessary in some applications, the finger motion still play an important role in manipulating gestures which are indispensable in applications such as interacting with virtual objects [11]. Due to self-occlusion, some features may not be available which makes the estimation difficult. At the same time, feature extracting itself may not be accurate and reliable. Robust real-time tracking and integrating multiple cues need further research.

Realistic hand animation should be considered in the future. Schemes of avoiding the violation of body constraints and collision detection should be built in animation systems. Combing hand gesture and speech should also be adequately addressed in the future to make a natural HCI [11]. Two-handed gestures should also be studied in the future.

6 Conclusions

In this paper, we report the past development on the research of human hand modeling, analysis and animation in the context of HCI. Hand can be modeled with several aspects such as shape, kinematical structure and dynamics. Different hand models are used in different applications. The 3D hand models offer a rich description to fully capture hand motion. Realistic hand animation can be achieved by motion editing techniques.

Overall, human hand modeling, analysis and animation are still in their infancy at the current state of the art. In order to develop a natural and reliable hand gesture interface, substantial research effort in computer vision, graphics, machine learning and psychology will be needed.

Acknowledgment

This work was supported in part by National Science Foundation Grant IRI-9634618 and Grant CDA-9624396, and in part by Army Research Laboratory Cooperative Agreement No. DAAL01-96-2-0003.

References

[1] J.K.Aggarwal, Q.Cai, "Human Motion Analysis: A Review, " *IEEE proc. Nonrigid and Articulated Motion Workshop'97*, pp90-102, 1997.

[2] S.Ahmad, "A Usable Real-Time 3D Hand Tracker", *IEEE Asilomar Conf.*, 1994

[3] James Davis, Mubarak Shah, "Towards 3-D Gesture Recognition", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol.13 No.3, May 1999.

[4] D.M.Gavrila, "The Visual Analysis of Human Movement: A Survey", *Computer Vision and Image Understanding*, Vol.73, No.1, Jan, pp.82-98, 1999

[5] S.Gortler, et.al., "Efficient Generation of Motion Transitions using Spacetime Constraints", *SIGGRAPH*, 1996

[6] T.Heap, D.Hogg, "Towards 3D Hand Tracking Using a Deformable Model", *Proc.Int'l Conf.Automatic Face and Gesture Recognition*, Killington, Vt., pp.140-145, Oct.1996

[7] K.H.Jo, Y.Kuno and Y.Shirai, "Manipulative Hand Gesture Recognition Using Task Knowledge for Human Computer Interaction", *Proc. 3rd IEEE International Conference on Face and Gesture Recognition*, pp.468-473, 1998.

[8] R.Kjeldesn, J.Kender, "Toward the Use of Gesture in traditional User Interfaces", *IEEE Automatic Face and Gesture Recognition*, pp151-156, 1996

[9] J.J.Kuch, T.S.Huang "Vision-Based Hand Modeling and Gesture Recognition for Human Computer Interaction" *Master thesis, Univ.of Illinois at Urbana-Champaign*, 1994.

[10] J.Lee, T. Kunii, "Model-based Analysis of Hand Posture", *IEEE Computer Graphics and Applications*, Sept, pp.77-86,1995.

[11] Vladimir Pavlovic, Gregory Berry, and Thomas Huang, "A Multimodal Human-Computer Interface for Control of a Virtual Enviornment", *American Association for Artificial Intelligence 1998 Spring Symposium on Intelligent Environments*, 1998

[12] Vladimir I. Pavlovic, R.Sharma, T.S.Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review," *IEEE PAMI*, Vol.19, No.7, July, pp.677-695, 1997.

[13] J.Rehg, T.Kanade, "Model-Based Tracking of Self-Occluding Articulated Objects" *Proc. of the IEEE Int'l Conf. Computer Vision*, Cambridge, Mass. pp.612-617, June 20-23 1995.

[14] N.Shimada, et.al., "Hand gesture estimation and model refinement using monocular camera - Ambiguity limitation by inequality constraints ", *Proc. of The 3rd Conf. on Face and Gesture Recognition*, pp.268–273,1998

[15] T.Starner, et.al. "A Wearable Computer Based American Sign Language Recognizer", *IEEE Int'l Symposium on Wearable Computing*, Oct, 1997

[16] J.Triesch, C.Von der Malsburg, "A Gesture Interface for Human-Robot-interaction", *IEEE Automatic Face and Gesture Recognition*, pp546-551, 1998

[17] L.Tsap, et.al., "Human Skin and Hand Motion Analysis from Range Image Sequences Using Nonlinear FEM", *IEEE Proc. Nonrigid and Articulated Motion Workshop*, pp.80-88, 1997

[18] C.Vogler and D.Metaxas, "ASL recognition based on a coupling between HMMs and 3D motion analysis", *Proc. of the Int'l Conf. on Computer Vision*, pp. 363-369, India, January 4-7, 1998.

[19] A.Wilson and A.Bobick, "Recognition and Interpretation of Parametric Gesture", *Proc. of the Int'l Conf. on Computer Vision*, 1998

[20] Ying Wu, Thomas S. Huang, "Capturing Articulated Human Hand Motion: A Divide-and-Conquer Approach", *Proc. of the IEEE Int'l Conf. on Computer Vision*, 1999.

[21] M. Zhao and F. Quek, "RIEVL: Recursive induction learning in hand gesture recognition", *IEEE Transactions on Pattern Recognition and Machine Intelligence*, Vol. 20, No. 11, November 1998, pp. 1174–1185