Determining 3-D Hand Motion*

James Davis†

Media Lab
Massachusetts Institute of Technology
Cambridge, MA 02139

Mubarak Shah

Computer Vision Lab
University of Central Florida
Orlando, FL 32826

Abstract

This paper presents a glove-free method for tracking hand movements using a set of 3-D models. In this approach, the hand is represented by five cylindrical models which are fit to the third phalangeal segments of the fingers. Six 3-D motion parameters for each model are calculated that correspond to the movement of the fingertips in the image plane. Trajectories of the moving models are then established to show the 3-D nature of hand motion.

1 Introduction

The importance of human gestures has been greatly underestimated. We use hundreds of expressive movements every day [2, 9], with many of these movements pertaining to hand gestures. These movements may have radically different interpretations from country to country – one hand gesture may represent a meaning of "good" in one country, whereas in another country it may be viewed as offensive [9]. Finger-spelling, a subset of sign language, permits any letter of the English alphabet to be presented using a distinct hand gesture. Using the finger-spelling gesture set, people can communicate words to one another using only hand movements [4]. The media has realized the significance of gestures and was experienced in the final scene of the movie, *Close Encounters of the Third Kind* (**Columbia Pictures**, 1977), where a human and alien communicated to each another using hand movements. **McDonald's** demonstrated the utilization of gestures in a 1994 television commercial showcasing patrons ordering any one of four different meals using the appropriate hand gesture. If we are to enhance and extend the man-machine interface, it is imperative to enable computers to interpret hand movements and to act intelligently according to their meanings.

Tracking hand motion becomes more realistic with a 3-D, rather than a 2-D, approach. With 3-D information, we know the real-world location of the fingers at any time, and can exploit this knowledge to suit applications without having to concern ourselves with the weaker and possibly ambiguous 2-D information. Two-dimensional ambiguities which may arise are the 3-D trajectories which, after undergoing perspective projection, have the same corresponding 2-D trajectory. Also, using 3-D models and motion parameters avoids the need for motion correspondence for mapping feature points to their correct 2-D trajectory [10, 6], for each feature point is a member of a distinct model for a particular finger and thus has no ambiguity in which trajectory it belongs. Therefore to remove these uncertainties which may arise in 2-D, we can use 3-D information.

In this paper, we discuss our method for developing a computer vision system which has the ability to model and track rigid 3-D finger movement of a glove-free hand. In the rest of this paper we first identify the fingers in the image (Section 3.1) and fit a 3-D generalized cylinder to the third phalangeal segment of each finger (Section 3.2). Then six 3-D motion parameters are calculated for each model corresponding to the 2-D movement of the fingers in the image plane (Section 4). Experiments are shown with 3-D hand movements (Section 5).

2 Related Work

Regh and Kanade [11] describe a model-based hand tracking system called *DigitEyes*. This system uses stereo cameras and special real-time image processing

*The research reported here was supported by the National Science Foundation grants CDA-9200369, IRI 9122006, and IRI 9220768.

†The author was a member of the Computer Vision Lab at the University of Central Florida during this research.

hardware to recover the state of a hand model with 27 spatial degrees of freedom. In order for *DigitEyes* to be used in specific hand applications, the kinematics, geometry, and initial configuration of the hand must be known in advance. Hand features are measured using local image-based trackers within manually selected search windows. Rendered models and state trajectories are given demonstrating the 3-D nature of their results.

Darrell and Pentland [5] have proposed an approach for gesture recognition using sets of 2-D view models of a hand (one or more example views of a hand). These models are matched to stored gesture patterns using dynamic time-warping, where each gesture is warped to make it of the same length as the longest model. Matching is based upon the normalized correlation between the image and the set of 2-D view models. This method requires the use of special-purpose hardware to achieve real-time performance, and uses gray-level correlation which can be highly sensitive to noise.

Cipolla, Okamoto, and Kuno [3] present a structure from motion (SFM) method in which the 3-D visual interpretation of hand movements is used in a man-machine interface. A glove with colored markers is used as input to the vision system and movement of the hand results in motion between the markers in the images. The authors use the affine transformation of an arbitrary triangle formed by the markers to determine the projection of the axis of rotation, change in scale, and cyclotorsion. This information is used to alter the position and orientation of an object displayed on a computer graphics system.

Segan's [12] *Gest* is a computer vision system that learns to identify non-rigid 2-D hand shapes and computes their pose. The system displays a hand in a fixed position on the screen and the user responds by presenting that same gesture to the camera. The hand's pose is calculated and classified. Recognition involves graph matching and employs a preclassifier to offset the matching cost. Each gesture is determined from the hand's 2-D position, and does not use any motion characteristics or 3-D feature locations. *Gest* was used to control graphics applications, such as a graphics editor and flight simulator.

Kang and Ikeuchi [8] describe a framework for determining 3-D hand grasps. An intensity image is used for the identification and localization of the fingers using curvature analysis, and a range image is used for 3-D cylindrical fitting of the fingers. A *contact web* is used to map a low-level hand configuration to a more abstract grasp description. The grasp is then identified using a *grasp cohesive index*. Though this method uses 3-D finger information, it requires both intensity and costly range imagery to produce the finger models.

In an earlier paper [6], we presented a method for recognizing hand gestures using a 2-D approach. A finite state machine is used to model four qualitatiely distinct phases of a generic gesture. If the hand is found to be in motion to the gesture position, fingertip trajectories are created using motion correspondence of the fingertip points in the image plane. Vectors are then used to approximate the trajectories, and the unknown gesture is matched to a library gesture using these vectors. Results show recognition of seven gestures (representatives for actions of *Left*, *Right*, *Up*, *Down*, *Grab*, *Rotate*, and *Stop*) without the use of any special hardware.

3 Finger Modelling

To generate an appropriate 3-D model for the hand, we require only one intensity image of the user's hand in a predefined start position. To begin, we first identify the fingers within the image and determine each finger's axis of orientation. Then generalized cylinders are fit to specific finger segments. Anatomical knowledge of the human hand is exploited to enhance the modelling process.

3.1 Identification of Finger Regions

Initially, we constrain the user to begin with the hand in a known start position (See Fig. 1.a). Using histogram thresholding, the original image is converted into a binary image in which small regions are removed (See Fig.1.b). We then find a set of points which can be used to differentiate the fingers from the rest of the image. Previous approaches for finding feature points involve boundary curvature extrema [8], interest operators to detect specially colored regions [3], and manual selection [11]. Our approach uses the knowledge of the start position and natural design of the hand to automatically determine five fingertip points $\{T_n\}_{n=0}^{4}$ and seven base points $\{B_m\}_{m=0}^{6}$ which are used to segment the fingers. Each finger region is found by applying a connected component algorithm using the respective fingertip and base points as bounds in the segmentation (See Fig. 1.c). Once the fingers have been identified, the axis of orientation for each finger can be calculated (See Fig. 1.d). The orientation axis is established by finding the line in which the integral of the square of the distance to

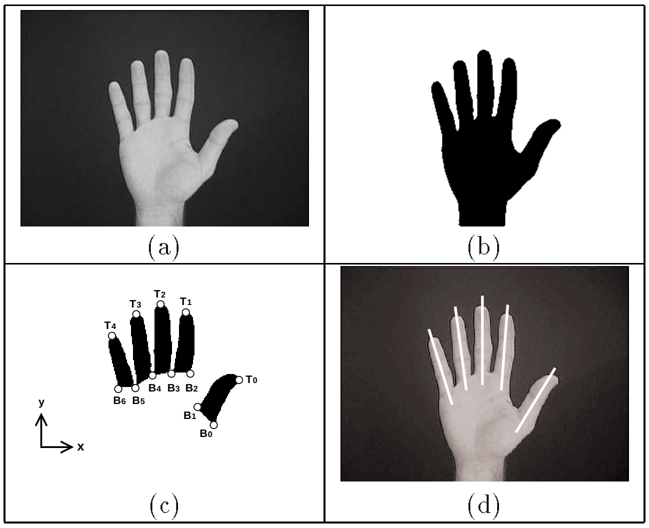

Figure 1: Determining Finger Orientation. (a) Start position of the hand. (b) Binary image resulting from histogram thresholding and removal of small regions. (c) Finger regions found using fingertip points $\{T_n\}_{n=0}^{4}$ and base points $\{B_m\}_{m=0}^{6}$. (d) Start position showing each finger's orientation axis.

points in the finger is a minimum. The integral to be minimized over finger F is

$$E = \int \int_F r^2 \; dx \; dy \quad , \tag{1}$$

where r is the perpendicular distance from point (x, y) to the axis sought after [7]. The fingers and axes will be used in generating cylindrical representations of finger segments.

3.2 Cylindrical Fitting

Cylindrical models can be employed to represent the fingers due to the inherent cylindrical nature of fingers. A finger as a whole is a non-rigid object, with the first phalangeal (FP), second phalangeal (SP), and third phalangeal (TP) segments (only FP and TP segments for thumb) [13] each exhibiting rigid behavior. We dismiss the concerns for non-rigidness, occlusion, and connectedness, and only model and track the TP segments (fingertip segments) for simplicity. To model the TP segments, we must know where they are located with respect to each finger in the image. In general, each FP, SP, and TP segment length occupies nearly a third of the total finger length. Using this heuristic, the major axis for the finger can be divided

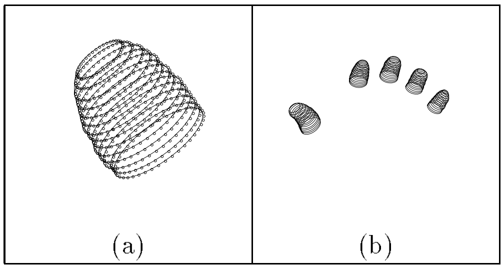

Figure 2: TP Models. (a) Index finger's 3-D cylindrical TP model shown with nodes. (b) All five TP models representing a model set for the hand.

into three parts (except for the thumb, where it is divided into two), designating the TP segment as the upper most third of the finger (upper half for the thumb) along the axis of orientation. A right straight homogeneous generalized cylinder (RSHGC)[14] can then be fit to give a 3-D model to each 2-D TP segment (See Fig. 2.a&b), such that each model's projection conforms to the actual respective fingertip in the image. A cross-section shape of an ellipse is used to fit the natural cross-section of a finger, with semi-major axis a and semi-minor axis b, having $b = f(a) \mid f(a) < a$.

4 Motion Parameter Estimation

Given a set of TP models and a sequence of intensity images in which the hand is moving, we would like compute the 3-D motion of the fingertips employing the 2-D motion in the image plane. The 3-D motion of a model is represented in terms of translation (T_x, T_y, T_z) and counter-clockwise rotation $(\omega_x, \omega_y, \omega_z)$ around the three coordinate axes based at the model's centroid. Our approach incorporates a direct method using spatio-temporal derivatives (instead of optical flow), a linearized rotation matrix (due to small motion changes), and a 3-D model (where the depth is known) to compute the 3-D motion. An over constrained set of equations is established and solved for the unknown motion parameters. The locations of the TP models are continually updated in 3-D to match the 2-D fingertip movement. Only visible model nodes can be used in the motion parameter calculation and can be determined by using together two methods (surface normals and depth array) for back-side elimination [1]. This process must be performed each time the model location is updated to ensure that previously visible nodes have not become occluded and vice-versa.

4.1 Formulation of Motion Parameter Estimation

Consider the optical flow constraint equation:

$$f_x u + f_y v + f_t = 0 \ , \tag{2}$$

where $f_x = \frac{\partial f}{\partial x}$, $f_y = \frac{\partial f}{\partial y}$, $f_t = \frac{\partial f}{\partial t}$, $u = \frac{dx}{dt}$, and $v = \frac{dy}{dt}$. Assume that the geometry projection from 3-D space onto the 2-D image plane is perspective projection with camera focal length F. Then the optical flow field (u, v) induced by the 3-D instantaneous motion about the object centroid is given by:

$$u = \frac{F}{Z}\left[(T_x + \omega_y Z_c - \omega_z Y_c) + \frac{-X}{Z}(T_z + \omega_x Y_c - \omega_y X_c)\right] \ , \tag{3}$$

$$v = \frac{F}{Z}\left[(T_y + \omega_z X_c - \omega_x Z_c) + \frac{-Y}{Z}(T_z + \omega_x Y_c - \omega_y X_c)\right] \ , \tag{4}$$

where (T_x, T_y, T_z) is the forward translation vector, $(\omega_x, \omega_y, \omega_z)$ is the counter-clockwise rotation vector, (X, Y, Z) are the world coordinates, and (X_c, Y_c, Z_c) are the object centered coordinates.

Substituting the above equations for u and v in (2) and rearranging, we get

$$\begin{aligned} -f_t &= f_x \frac{F}{Z}\left[(T_x + \omega_y Z_c - \omega_z Y_c) + \frac{-X}{Z}(T_z + \omega_x Y_c - \omega_y X_c)\right] \\ &+ f_y \frac{F}{Z}\left[(T_y + \omega_z X_c - \omega_x Z_c) + \frac{-Y}{Z}(T_z + \omega_x Y_c - \omega_y X_c)\right] \end{aligned} \tag{5}$$

which can also be written as

$$\begin{aligned} -f_t &= \left[f_x \frac{F}{Z}\right] T_x + \left[f_y \frac{F}{Z}\right] T_y - \left[\frac{F}{Z^2}(f_x X + f_y Y)\right] T_z \\ &- \left[\frac{F}{Z^2}(f_x X Y_c + f_y Z Z_c + f_y Y Y_c)\right] \omega_x \\ &+ \left[\frac{F}{Z^2}(f_x Z Z_c + f_x X X_c + f_y Y X_c)\right] \omega_y \\ &- \left[\frac{F}{Z}(f_x Y_c - f_y X_c)\right] \omega_z \ . \end{aligned} \tag{6}$$

In this equation, (X, Y, Z) and (X_c, Y_c, Z_c) are known from the model, and f_x, f_y, and f_t can be computed from image pairs. Therefore the only unknowns are the motion parameters (T_x, T_y, T_z) and $(\omega_x, \omega_y, \omega_z)$. An over constrained set of equations is established using visible nodes and in matrix form is as follows

$$[\mathbf{A}]\,\mathbf{x} = \mathbf{b} \ ,$$

with $\mathbf{x} = (T_x, T_y, T_z, \omega_x, \omega_y, \omega_z)^T$. A linear regression using least squares is used to approximate the six unknown motion parameters in $\mathbf{x}$, and is iterated to account for linearizing. Initially, for calculating the motion parameters between frame 1 and frame 2, the visible model nodes record the corresponding intensity and gradient information from frame 1. Then the motion parameters are determined using the model nodes and frame 2. After application of the parameters to the model from frame 1, the model is now located to conform to frame 2. For frame 3, a new estimation is calculated using the model (compensated from frame 1 to frame 2) and frame 3. This process, continues for the remainder of the sequence.

5 Experiments

Our system was used to track two distinct hand motions: movement in the XY plane (See Fig. 3 Sequence 1), and movement in the XZ plane, i.e. scaling (See Fig. 3 Sequence 2). These examples are sufficient to demonstrate the advantage of a 3-D, rather than a 2-D, approach. In each sequence, the locations of the TP models were updated in each frame to match the movement of the fingertips in the image plane (See superimposed models in Fig. 3.1a&b and Fig. 3.2a&b). In sequence 1, with no depth changes, the 2-D trajectories are shown to be adequate to approximate the motion of the hand (Compare 2-D and 3-D trajectories in Fig. 3.1c&d). Sequence 2 demonstrates the hand changing in depth. This type of motion can be shown in 3-D (See 3-D trajectories in Fig. 3.2d) and cannot be distinguished in 2-D, where it appears that the hand is mainly at rest (See 2-D trajectories in Fig. 3.2c).

6 Conclusion

In this paper, we presented a 3-D hand modelling and motion estimation method for tracking hand movements. This approach does not require any glove or motion correspondence, and recovers 3-D motion information of the hand. The orientation of the fingers in a 2-D image are found, and a generalized cylinder is fit to each finger's third phalangeal segment. Six motion parameters for each finger are calculated, which correspond to the 2-D movement of the fingertips in the image plane. Three-dimensional trajectories are then determined from the motion of the models, which may be used in hand tracking and gesture recognition applications.

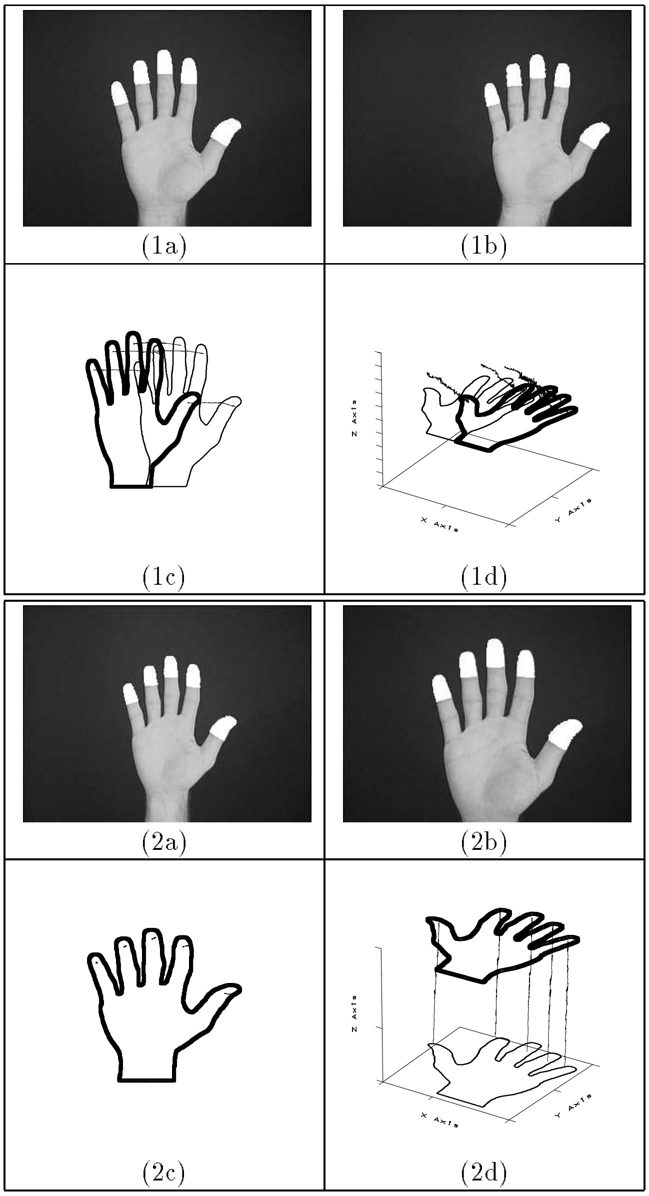

Figure 3: Sequence 1: XY translation. (1a)&(1b) First and last images with TP models (white). (1c) 2-D motion trajectories. (1d) 3-D motion trajectories. Sequence 2: XZ translation. (2a)&(2b) First and last images with TP models (white). (2c) 2-D motion trajectories. (2d) 3-D motion trajectories. (Note: bold hand outline represents initial hand position.)

Acknowledgment

We would like to acknowledge Reinhold Koch and Niels da Vitoria Lobo for their suggestions.

References

[1] Artwick, B. *Applied Concepts in Microcomputer Graphics.* Prentice-Hall, New Jersey, 1984.

[2] Bauml, B., and Bauml, F. *A Dictionary of Gestures.* The Scarecrow Press, New Jersey, 1975.

[3] Cipolla, R., Okamoto, Y., and Kuno, Y. Robust structure from motion using motion parallax. In *ICCV*, pages 374–382. IEEE, 1993.

[4] E. Costello. *Signing: How to Speak With Your Hands.* Bantam Books, New York, 1983.

[5] Darrell, T., and Pentland, A. Space-time gestures. In *CVPR*, pages 335–340. IEEE, 1993.

[6] Davis, J., and Shah, M. Recognizing hand gestures. In *ECCV*, pages 331–340, May 1994.

[7] Horn, B.K.P. *Robot Vision.* McGraw-Hill, 1986.

[8] Kang, S.B., and Ikeuchi, K. Toward automatic robot instruction from perception – recognizing a grasp from observation. *IEEE Transactions of Robotics and Automation*, 9:432–443, August 1993.

[9] Morris, D., Collet, P., Marsh, P., and O'Saughnessy, M. *Gestures: Their Origins and Distribution.* Stein and Day, 1979.

[10] Rangarajan, K., and Shah, M. Establishing motion correspondence. *CVGIP: Image Understanding*, 54:56–73, July 1991.

[11] Rehg, J., and Kanade, T. Visual tracking of high dof articulated structures: an application to human hand tracking. In *ECCV*, pages 35–46, May 1994.

[12] Segen, J. Gest: A learning computer vision system that recognizes hand gestures. *Machine Learning IV*, 1994.

[13] Taylor, C., and Schwarz, R. The anatomy and mechanics of the human hand. *Artificial Limbs*, 1955.

[14] Zerroug, M., and Nevatia, R. Segmentation and recovery of shgcs from a real intensity image. In *ECCV*, 1994.