Virtual Kathakali: Gesture-Driven Metamorphosis

Soumyadeep Paul, Sudipta N. Sinha, and Amitabha Mukerjee
Center for Robotics
Indian Institute of Technology, Kanpur
e-mail: {spaul,snsinha,amit}@iitk.ac.in

Abstract

Training in motor skills such as athletics, dance, or gymnastics is not possible today except in the direct presence of the coach/instructor. This paper describes a computer vision based gesture recognition system which is used to metamorphose the user into a Virtual person, e.g. as a Kathakali dancer, which is graphically recreated at a near or distant location. Thus this can be seen by an off-site coach using low-bandwidth joint-motion data which permits real-time animation. The metamorphosis involves altering the appearance and identity of the user and also creating a specific environment possibly in interaction with other virtual creatures. Unlike previous approaches to gesture based virtual reality, here i) a single monochrome camera is used to track the arms, and ii) colour differences are used to disambiguate situations where the arm is occluding the user's body.

A robust vision module is used to identify the user, based on very simple binary image processing in real time which also manages to resolve self occlusion, correct for clothing/colour and other variations among users. Gestures are identified by locating key points at the shoulder, elbow and wrist joint, which are then recreated in an articulated humanoid model, which in this instance, represents a Kathakali dancer in elaborate traditional dress. Unlike glove based or other gesture and movement tracking systems, this application requires the user to wear no hardware devices and is aimed at making gesture tracking simpler, cheaper, and more user friendly.

1 Introduction

Recognizing and tracking human gestures, with its great promise for simplifying the man-machine interface, has seen considerable emphasis in recent years. The main approaches have been to use dedicated hardware such as dataglove or polhemus sensors, [10, 14] or visual recognition which requires little hardware but yields less direct results [1, 2, 4, 5, 6, 9, 11, 12, 13, 15, 16].

One application of gesture recognition that is beginning to emerge is off-site training for motor skills, e.g. in activities such as athletics, surgery, theater/dancing, or gymnastics. Here the user's motions can be transmitted using some low-bandwidth representation (e.g. joint angles or facial expressions), and the instructor or coach at the remote site can visualize the student using a local graphics model at real-time animation rate, and provide appropriate feedback, possibly illustrating the correct procedure through a similar virtual metamorphosis channel. For example, a renowned master in the Kathakali dance form [1] may be able to provide personalized feedback to a student far away - the sensation of being co-located in the same virtual space makes communication much more natural. Furthermore, the master (or disciple) has the ability to zoom in on a particular part of the performance or view the scene from a particular vantage point, or to have the actions repeated in slower speeds.

Of course, other usual Virtual Reality applications such as full body interaction in a virtual space, as in games or advanced chat rooms, can also be conducted with such a system. Figure 1 shows the basic setup that would be needed.

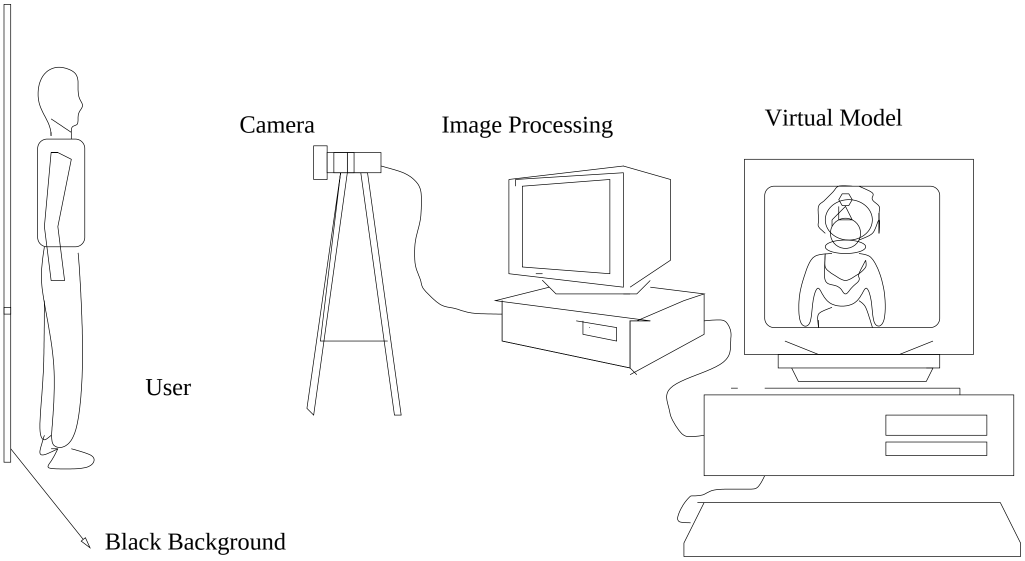

Figure 1: *Virtual Modeling of User Motions.* The user moving in front of the camera sees himself as a Kathakali dancer in the Virtual Model. The gestures of the user are transmitted to an articulated graphics model which then recreates it with the appropriate costume and other embellishments. This low-bandwidth data may be used to create an off-site display for obtaining the trainer's feedback.

[1] *Kathakali:* This celebrated dance tradition involves wearing elaborate costumes and headgear, and also the use of special eye-masking paints and other cosmetics.

Early approaches to gesture modeling used specialized arm motion detection sensors [14]. Such sensors encumber the user and impose constraints on their motion to a certain extent. The camera based model provides a simpler, more flexible, and far cheaper alternative to other approaches. However, with the camera the body pose is not directly available, and considerable effort is needed in image processing. Different parts of this problem have been tackled for many years now:

- The DigitEyes system Rehg/Kanade 94 ([12]) recovered a detailed kinematic description of the hand using a 27 degrees of freedom full hand model and one or two cameras.
- Markov model based gesture identification [1, 13].
- Body tracking and behaviour interpretation [16].

In general, camera based systems are not able to simultaneously identify both fine and gross motions since a full body field of view reduces the accuracy available for looking at the hand. See [9] for a recent survey of the field.

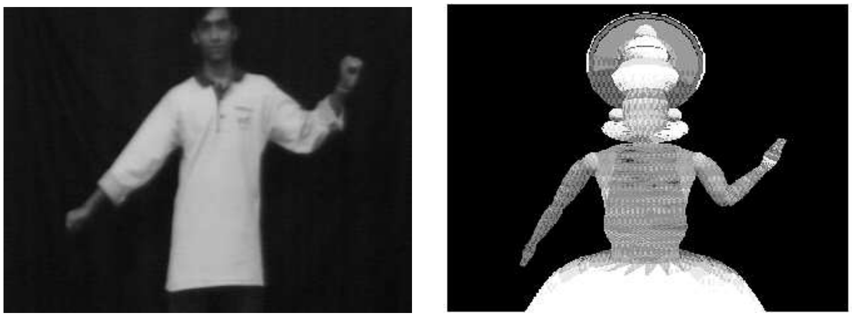

Figure 2: *The User and the Model.* The User pose as seen by the camera (from which the arm pose is to be detected), and the final Kathakali Dancer Model as displayed to the user.

Combining gesture recognition with graphics reconstruction provides a virtual space where the user's action can be reflected. Applications in this genre include games [4, 6], Virtual interaction spaces [3], remote tele-operation [5], and Virtual Metamorphosis [8]. Our application is in the metamorphosis category where the user is metamorphosed as a Kathakali dancer in a virtual environment. The following section gives a brief outline of the techniques used in the paper.

2 Outline

The system can be broken up into three modules:

- Real-time Detection of arm movements of the user (Section 3).
- Modelling of the Kathakali dancer (Section 4).
- Reproduction of the pose in the Kathakali dancer's model (Section 5).

Unlike other models that use thermal imaging to obtain the user's silhouette [7], the Virtual Kathakali system uses a visible-light monochrome camera against a black background. The user's silhouette is obtained by dynamically binarizing the images and the 3-D positions of the user's shoulder, elbow and the wrists are obtained in real time from the image coordinates. This compact data is then transmitted to a local or off-site virtualization system in real time. Also, by using skin-tone colour/greyscale information, it is possible to identify occlusion, which is not possible to do in thermal imaging systems. The overall cost of this system is likely to be several times less than that of other comparable systems used in Virtual Metamorphosis systems.

The next phase is to create an articulated 3D model that will follow the user's poses and reflect the traditional costumes of a classical Indian dance form such as the Kathakali. The 3D model needs to have appropriate motion constraints at the joints and suitable dress/headgear/texture. The 3D arm pose sequence is now communicated to the graphics model, which recreates it as an animated graphics display. In this process, the very low-bandwidth joint angle data can be used to animate the 3D Dancer model.

3 Real Time detection of User's Arm Pose

In the initial pose, the user stands with his arms separated wide apart and the following calibration data are obtained :-

- arm-length
- range of pixel intensity within which the pixels corresponding to his body tone lie
- arm width
- width of his shoulders

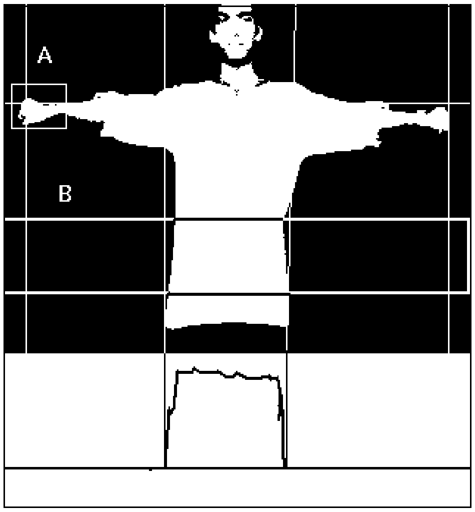

Figure 3: *Calibration phase.* User stands with arms apart. The body width, arm length and arm width are detected. The histogram for the box B (shown below the image) is used to detect body width, and a similar analysis in the region A is used to obtain the skin shade.

To simplify the image processing costs, the user is required to stand in front of a dark background. This permits image binarization based on a dynamic threshold, set at a point of sharp variation in the intensity histogram. Noise may yet interfere with the robust determination of arm posture, and Gaussian convolution is used to smooth out some of the noise.

The pose of the user's arms at each instant is obtained by identifying the elbow and wrist in the image. The body width is identified based on points of high intensity change in the lower parts of the image. The elbow and wrist are distinguished by different techniques depending on whether the hand is occluding the body (Section 3.1) or not. In the latter case, a fast and simple technique is to locate the extreme points in the image and test if the line joining it to the shoulder is part of the arm or not (Figure 5). Based on this and the initial calibration information, the 3D pose of the arm is estimated based on foreshortening.

3.1 Resolving Occlusion

Many dance poses involve the hand being in front of the body; these postures are particularly important in many *mudras*[2]. The binarized processing described

[2] *mudras:* certain specific postures or positions in Indian dance forms.

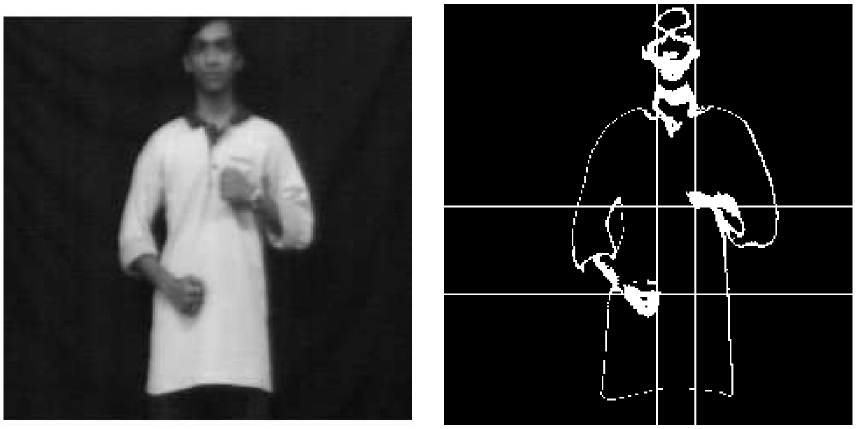

Figure 4: *Resolving Occlusion.* The pixel intensity range for the user's skin colour / clothes are used to distinguish parts of the arm that may be occluding the user's body.

above is not sufficient for resolving this occlusion, so the wrist is identified in the image based on the greyscale skin tones (see Figure 3). During the performance, for each grabbed image, the pixels outside this intensity range are discarded. If the user is wearing clothes contrasting with his body color, only the pixels corresponding to his body parts retain the high value of intensity. This leads to an effective separation of the hand from the body in cases of self-occlusion (see Figure 4). The hand can now easily be detected.

In the case where the hand is before the body, we still need to identify the elbow for constructing the 3-D arm pose. This is simple here, since the outermost tip of the image corresponds to the elbow joint.

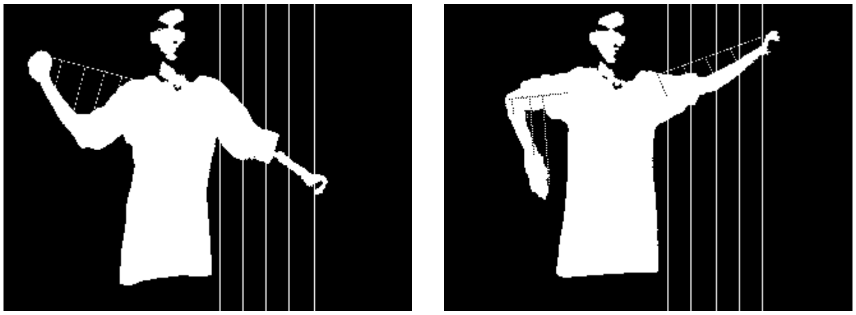

Figure 5: *Image Processing Results.* Left: Locating outermost tip. Right: Finding the elbow and wrist joints.

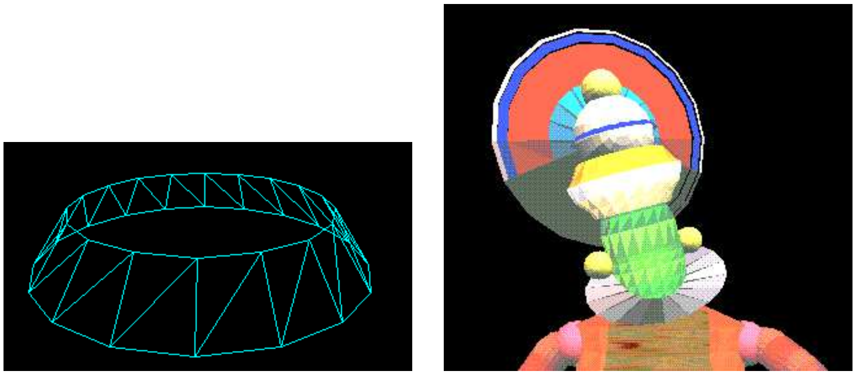

Figure 6: *Graphics Modeling.* The frustum primitive has been used in several constructions, for example the headgear shown to the right.

4 Modelling of Kathakali Dancer

The first step is to create the wireframe of the human using primitives. A frustum with variable elliptical cross-sections has been used as a building block for modelling the Kathakali dancer. The outer surface of this primitive is realised by a triangular mesh. A stack of such primitives are used to model the head, arms and body separately. Though the resolution can be easily controlled, a very high resolution is sacrificed for faster rendering.

The head has three degrees of freedom i.e.. it can be twisted, bent forward and sideways. Moreover, the movements are limited in the range that is humanely possible. To model the arms, first the shoulder joints with three degrees of freedom have been created. Next, the upper arm, the lower arm and the wrist are created using the frustum primitive. The elbow and wrist joints are simulated as hinge joints i.e.. with only one degree of freedom.

The Kathakali dance form is famous for its elaborate apparel and ornamentation. Hence, to give a realistic effect, texture mapping has been used to create an appropriate pattern on the front of the dress. The image used for texture mapping has been obtained from images of Kathakali dancers.

Various other Graphics rendering techniques like Lighting and Shading have been used to create the effect of a 3D Virtual World. The real time nature of the application requires the use of special techniques for optimizing performance.

Label	*Joint description*	*Number of DOF*
A	Neck joint	3
B,C	Shoulder Joints	3
D,E	Elbow Joints	1
F,G	Wrist Joint	1

Table 1: *Modelling of Joints* This Table shows the degrees-of-freedom and constraints of the labelled joints in Figure 7.

Head	*Upper-arm*	*Forearm*	*Palm*	*Body,Dress*
400	350	250	100	350

Table 2: *Number of Polygons* This Table shows the number of polygons used to create the model of Figure 7.

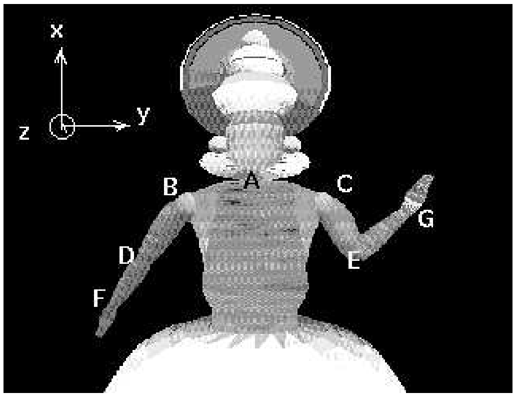

Figure 7: *The Articulated Model.* The human upper body is modelled as a set of rigid links connected by joints with different ranges of motion. The labels for the different joints are shown here (see Table 1).

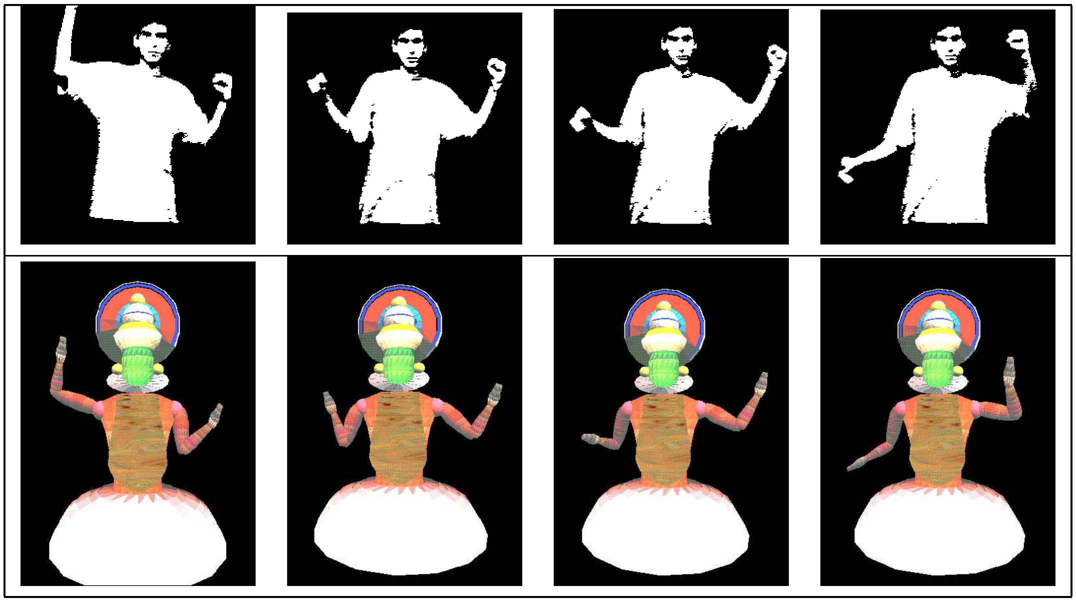

Figure 8: *User's and the Model's poses.* Four frames representing a motion sequence are shown along with the graphics reconstruction (lower row). For each frame, twelve joint angles are passed from the image interpretation module to the graphic generation module.

5 Reproduction of User's Pose in Virtual Model

Based on the initial data about user's arm-length and forearm-length the 3D position can be recovered using the foreshortening seen in the image. The angles of the upper arm with respect to the shoulders and the forearm with respect to the upper arm are calculated and used to apply appropriate transformations of the limbs in the virtual metamorphosis model.

This application was developed on a a PentiumII/Linux-OpenGL PC for the graphics and an older 486 PC/DOS with a Matrox Image Processing card under Matrox Imaging Library(MIL 2.1).

Some images with user's poses and the corresponding graphics output are shown in Figure 8.

6 Conclusions

In this paper, we have presented a communication environment in which a person could transform himself into any character he desires in a virtual environment. In our case, the character was that of a Kathakali Dancer. Computer Vision techniques are used for passive detection of user's arm pose in real time. Some simple constraints on the user (that he stand in front of a dark background, and that his skin tone differs from his clothing) result in being able to use some extremely simple and fast detection methods based on dynamic threshold binarization. A single camera is used and the user does not need to wear any special devices. The 3-D model is instantiated during the calibration phase, and the user's motions are reproduced using various transformations on the dancer model. Our implementation showed good performance with a processing speed of 5 frames per seconds on a very old and simple image processing system.

The primary application for such a system would be in the coaching of motor skills with distant trainers and coaches. In the arts, such a system would permit quick informal recording of creative insights, which would otherwise require elaborate stage and makeup arrangements before being presented in the final form. In general, the expansion of Internet, even with low-bandwidth connectivity, will permit gestural interaction of this type as long as major computational tasks such as identification of limb postures and recreation/graphics output are performed on the local machines. In the coming decades, systems of this type are likely to have a profound effect on the declining trend in artistic traditions worldwide. Also, in sports as well as critical operations such as surgery, this could enable expert trainers to provide guidance to a far larger set of students than would be possible in a face-to-face mode.

Since the image processing is carried on an image in 2D, the system is not able to resolve between two arm poses with the same projection, as when the entire arm is on a horizontal plane. However, due to the very nature of this deficiency, it will not matter to the viewer so long as it is viewed from the same angle.

In this work, we have only used the arm pose of the user to control the virtual model. In traditional dance form, facial expressions and finger movements constitute an important component of the dancer's emotional expression (*abhinaya*). In the current phase, with a single camera of fixed resolution, this is not possible; in fact, no vision system today can model both the finger and the gross body motions. However, some beginnings have been made towards integrating face recognition by having a camera look down on the user's face from a fixture mounted on the head itself. With multiple cameras, one camera could be used for the full-body field of view, whereas one more other cameras could track the important aspects of the dance - particularly the two hands and also the face. Such a multi-scale imaging and tracking system would enable detailed reconstructions of the important aspects of the scene.

Moreover the system, which supports only a single user at present, can easily be extended to support multiple users all sharing the same virtual environment.

Hence creation of Virtual Theaters become a distinct possibility where different actors situated far away from each other can be actors in the same virtual theater. Since the bandwidth required by the system is very small, transmitting the data over the Internet is a feasible option. Another challenging problem is to have real and virtual actors share the same space in a seamless manner from the audience's viewpoint.

References

[1] E. Hunter, J Schlenzig, and Ramesh Jain. Posture estimation in reduced-model gesture input systems. In *Intl. Workshop on Applications of Face and Gesture Recognition*, [`http://vision.ucsd.edu/papers/zurich95.ps.gz`], 1995.

[2] D. Kortenkamp, E. Huber, and P. Bonasso. Recognising and interpreting gestures on a mobile robot. In *Thirteenth National Conference on Artificial Intelligence, AAAI-96*, August 1996.

[3] M. W. Krueger. *Artificial Reality II.* Addison Wesley, 1991.

[4] N.K. Mishra, M.P. Singh, T.V. Prasannaa, B.K. Birla, D. Vidhani, A.N. Lal, and A. Mukerjee. Experiments in gesture based user interfaces. In *Conference of the Computer Society of India (CSI-96)*, October 1996.

[5] A. Mukerjee and S. K. Dash. Off-site tele-operation using gestures. In M. Vidyasagar, editor, *Proceedings ISIRS-98*, Bangalore, January 1998. (See [S.K. Dash, 1998] "Gesture-basd Teleoperation", M.Tech. Thesis, IIT Kanpur Dept Mech Engg, April 1998 for more details).

[6] Amitabha Mukerjee, Debabrata Dash, Amit, Binny S. Gill, and Mukesh P. Singh. Looking beyond the mouse: Gesture-based internet games. In B.N. Jain and S. Kanungo, editors, *Proceedings SEARC-97*. Tata McGraw-Hill, New Delhi, December 4-6 1997.

[7] J. Ohya and K. Sengupta. Generating virtual environments for human communications - virtual metamorphosis system and novel view generation. In *Computer Vision for Virtual Reality Based Human Communication (CVVRHC)*, pages 43–50, January 1998.

[8] Jun Ohya, Kazuyuki Ebihara, Jun Kurumisawa, and Ryohei Nakatsu. Virtual kabuki theater: Towards the realisation of human metamorphosis systems. In *IEEE International Workshop on Robot and Human Communication*, 1996.

[9] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 19(7):677–695, July 1997.

[10] P.K. Pook and D. Ballard. Tele-assistance: Contextual guidance for autonomous manipulation. In *AAAI-94*, pages 1291–1296, {pook,dana}@cs.rochester.edu, 1994.

[11] Francis K.H. Quek. Towards a vision-based gesture interface. In *Virtual Reality Software and Technology, Proceedings of the VRST '94 Conference*, pages 17–31. World Scientific, Singapore, 1994.

[12] James M. Rehg and Takeo Kanade. Digiteyes: Vision-based human hand tracking. Technical Report CMU-CS-93-220, CMU-CS, `[ftp://reports.adm.cs.cmu.edu/usr/anon/1993/CMU-CS-93-220.ps.Z]`, 1994. (Extended version of ECCV-94 paper).

[13] Thad Starner and Alex Pentland. Real-time american sign language recognition from video using hidden markov models. Technical report, MIT - The Media Laboratory, 1995.

[14] D. J. Sturman and D. Zeltzer. A survey of glove-based input. *IEEE Computer Graphics & Applications*, pages 30–39, January 1994.

[15] A. Wilson and A. Bobick. Recognition and interpretation of parametric gestures. In *Proceedings of ICCV 98*, Mumbai, January 1998. Also available as MIT Media Lab TR-421.

[16] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 19:780–785, July 1997.