# Visually Guided Interaction and Animation

Alex Pentland, Stan Sclaroff, Trevor Darrell, Irfan Essa, Ali Azarbayejani, Thad Starner
Perceptual Computing Section, Room E15-387
The Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{sandy,stan,trevor,irfan,ajazar,thad}@media.mit.edu

**Abstract**

We survey research at the M.I.T. Media Laboratory concerned with accurately modeling, tracking, and interacting with people. Applications include computer animation, user interfaces, and video understanding.

## 1 Introduction

The goal of this paper is to present an overview of our research concerned with modeling, tracking, and interacting with people. We will make no attempt to fully explain the technical underpinnings of our work or provide full references to the literature; these details are available in referenced articles, and in any case would not fit within the allotted space.

Papers describing the work surveyed here can be obtained by anonymous FTP from the computer whitechapel.media.mit.edu. C code implementing some of these algorithms can also be found at this FTP site.

The plan for the first half of this paper is to illustrate the estimation techniques we have developed for accurate modeling and tracking of humans:

- We will start by illustrating the our Kalman filter techniques for accurately tracking people.
- We will then show how by including object dynamics in the estimation process, it is possible to obtain accurate computer models of non-rigid biological motion.
- If we further include models of muscle control in the estimation process, it becomes possible to obtain accurate estimates of the underlying muscle activations.

Modeling geometry, motion, and muscle control requires making a large number of accurate measurements of image motion. This is because the situations we wish to model have a large number of degrees of freedom. Consequently, such detailed modeling is not currently possible in real-time systems without special hardware.

However in many situations we know in advance that there are a relatively small set of predetermined actions, or setting a relatively small number of "control knobs," that we care about. This is particularly characteristic of *interactive* systems, which typically have a limited repertoire of behaviors and reactions. In such systems there are relatively few independent geometric parameters, each of which may have a large degree of temporal variation. We have found that it is generally possible to set these parameters using simple, real-time visual measurements, making possible real-time processing using only standard computer workstations. The second half of this paper will focus on the real-time, interactive systems we have built using this methodology.

## 2 Measurement from Images

Ideally, we would like to have a real-time system that can recover the shape and motion of objects in a robust and flexible manner. The desire for real-time operation and arbitrary-length sequences strongly favors recursive techniques that integrate information from each new frame with prior accumulated information. The requirement of robustness suggests techniques that explicitly account for measurement noise and modeling uncertainty. These are both strengths of the extended Kalman filter (EKF), which has been the subject of much structure and motion estimation research.

We will illustrate the power of the EKF approach by the example of using video analysis to track a persons's head more accurately than is possible by use of a Polhemus sensor. Additional detail can be found in references [1, 2, 3].

In this example, a person's head was tracked using both the vision algorithm and the Polhemus magnetic



1058-6393/95 $4.00 © 1995 IEEE

sensor simultaneously. Figure 3 shows the vision and Polhemus measurements after an *absolute orientation* was performed to align the estimates properly. The RMS difference in translation is 1.25*cm* and the RMS difference in rotation is 1.5°. This accuracy is equal to the observed accuracy of the Polhemus sensor, indicating that the vision estimate is somewhat more accurate than the Polhemus sensor.

## 3 Physical Modeling

The EKF can be made more accurate by adding a detailed *physical model* to help it average together measurements in a more physically-meaningful manner. Physical modeling is by now a well-known approach in the computer vision literature; we make use of a variant known as *modal analysis* [9, 13, 10, 12]. The advantage of using this method is that it allows stable, closed-form solutions, and permits much more efficient physical simulation. For additional details see references [9, 13, 10, 14, 6, 12].

Figure 2 illustrates an example of using this technique to track an articulated object. This figure shows three frames from a twelve image sequence of a well-known tin woodsman caught in the act of jumping. Despite the limited range of motion, this example is a difficult one because of pronounced highlights on thighs and other parts of the body. The estimated motions for this sequence are illustrated by the bottom row of Figure 2. As can be seen by comparing the 3-D motion of the model with that in the original image, the resulting tracking is reasonably accurate.

## 4 Control Models

By incorporating physical models into the EKF, we can obtain accurate estimates of even non-rigid motion. However the descriptions we obtain in this manner are still *passive*, that is, they do not tell us about muscles activations, but only about the resulting movement and deformation. To obtain estimates of the muscle activations, it is necessary to agument the EKF with a *control model*, as shown in Figure 3. In this figure, the lower loop is the EKF (including the physical model), and the upper loop is the *control loop*. The output of this second control estimation process is the muscle activations required to account for the observed motion and deformation [7].

We have applied this methodology to the problem of measuring facial muscle activations. Because facial motion is so complex, the input must be very detailed and dense. Consequently, we use as input pixel-by-pixel measurements of surface motion (called *optical flow*) as input measurements. These dense motion measurements are then coupled to a physically-

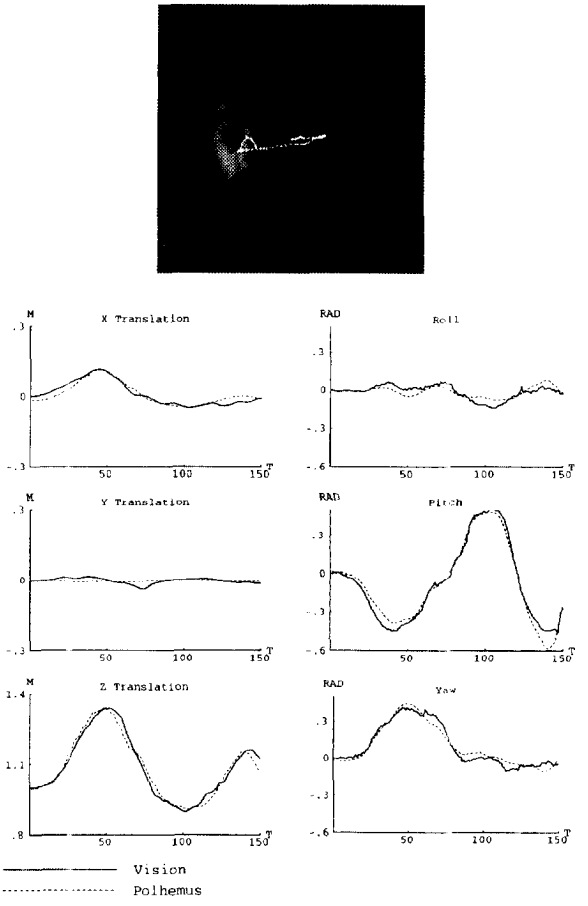


Fig. 1. Experiment 4: head tracking. Top: Head being tracked, Graphs: Vision and Polhemus estimates of head position. Much of the observed error is known to be due to Polhemus error. RMS differences are 1.25cm and 1.5 degrees.

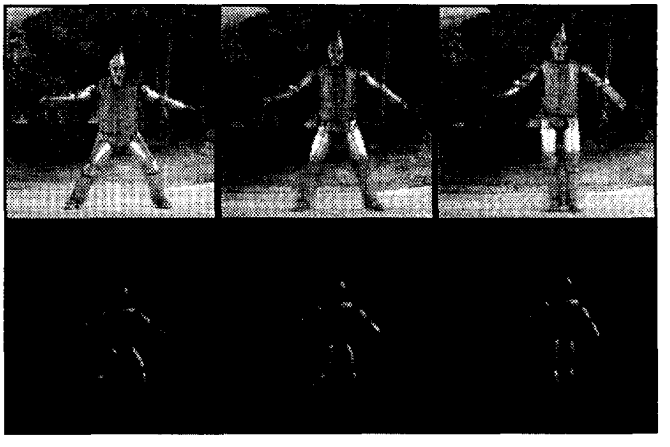

Fig. 2. Three frames from an image sequence showing tracking of a jumping man

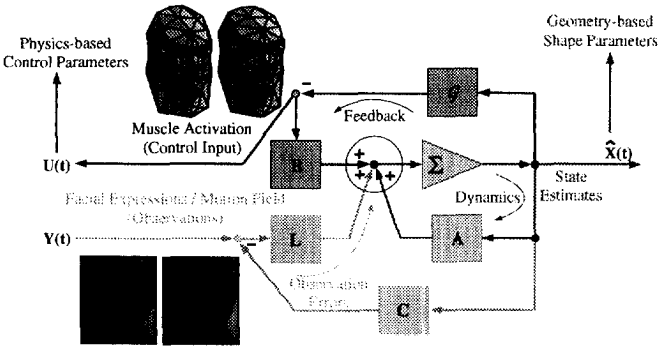


Fig. 3. *A block diagram of our modeling system, showing the estimation and correction steps, the dynamics loop, and the control feedback loop.*

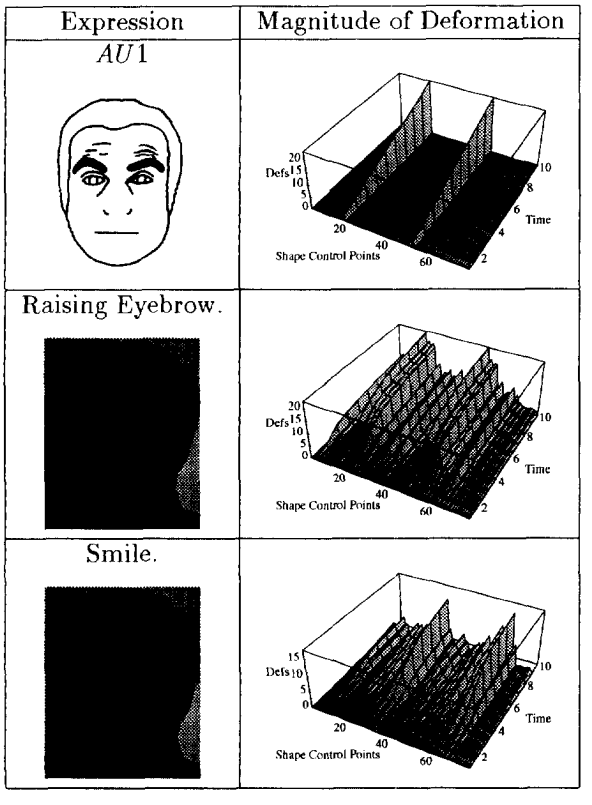


Fig. 4. *Observed Raising Brows and Smile versus static expression: Surface plots showing deformation over time for FACS actions AU1, and for an actual video sequence of raising eyebrows and smile.*

based face model *and* to a muscle control model. The outputs of this modeling process are detailed records of both the displacement of each point on the facial surface *and* the muscle control required to produce the observed facial motion. The recovered and muscle control patterns can be used to recognize facial expressions, animate other models, or composed to make new combination expressions.

The advantage of this approach over *a priori* facial modeling is that we can observe the complex muscle coarticulation patterns that are characteristic of real human expressions. For instance, it has been remarked that a major difference between real smiles and forced or faked smiles is motion near the corner of the eye. We have been able to observe and quantify the relative timing and amplitude of this near-eye motion using our system. For additional detail see reference [7].

## 5 Interactive Video Environements

We have constructed a number of real-time systems that we refer to as interactive video environments (IVE). The first of these systems was called ALIVE (Artificial Life Interactive Video Environment), which was demonstrated to over 500 participants at SIGGRAPH 1993 [8, 5]. This system used active, attention-driven vision to allow simulated "artificial life" agents to interact with real people through a video screen In this environment the agents and the user can "see" each other — users can see the agents on the video screen, and the agents can see users through a computer vision system. An image of the user appears on the video screen, effecting a type of "magic mirror", in which users see themselves in a different world through the use of a simulated mirror (Figure 5).

Developing vision routines for such an interactive systems offers several challenges relative to traditional image processing, and in particular calls for active/situated vision techniques. We have used a behavior-based approach to model both the action selection of the agents and their perception of the user. This has allowed us to use a large number of specialized image processing modules while still employing only limited processing power.

Recognizing and tracking hand gestures is a good example of specialized, attention-driving processing, because their appearance is very complex and because they are critical for natural interactions between human users and the computer agents [4].

We use a model of hand shape that is similar to the modal model referenced above; provides a coarse description of hand shape, position, orientation, and

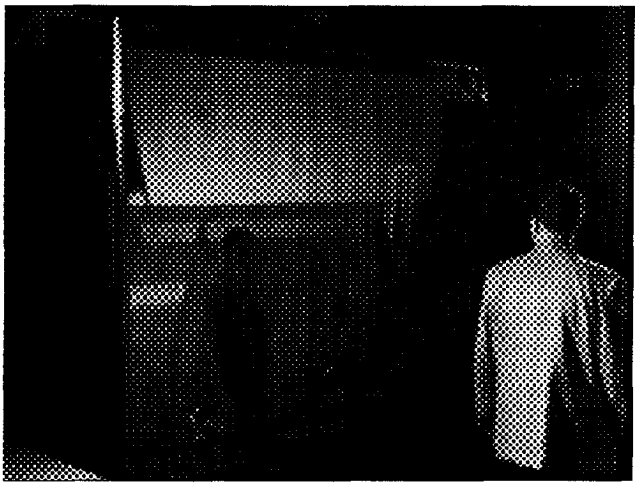

Fig. 5. The magic mirror metaphor: the user sees his/her mirror image surrounded by autonomous agents.

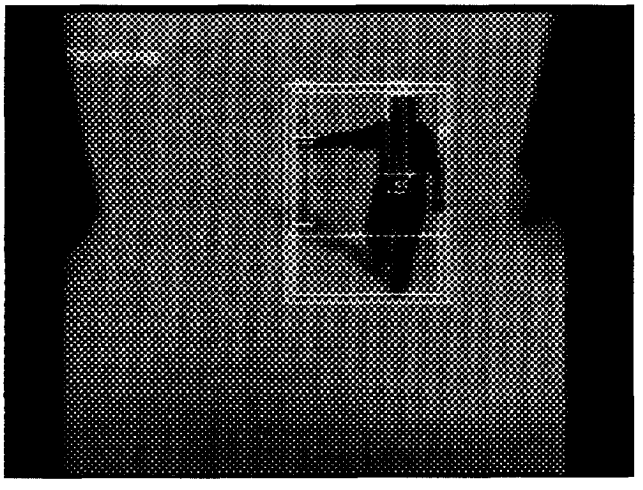

Fig. 6. The vision system works off a silhouette of the user. It computes a range of features including the bounding box, which is used to project the user's location in 3-D.

velocity. Each gesture is then defined in terms of these variables, and the temporal pattern that characterizes each gestures is learned by presenting many examples of each gesture (see Figure 6).

Using this training we are able to use temporal context to actively focus our visual routines, thus allowing us to recognize a large class of hand gestures. Recently we have used this approach to recognize hundreds of sentences in American Sign Language (ASL), using a base vocabulary of 40 word gestures [16].

## 6 Summary

We have developed a range of modeling, tracking, and interaction methods suitable for the human figure. Computer systems using these methods have allowed us to build a wide range of interactive video environment applications. Technical details are contained in the following references; these papers and software implementing some of these algorithms is available by anonymous FTP from the machine whitechapel.media.mit.edu.

## References


[1] Azarbayejani, A., Starner, T., Horowitz, B., and Pentland, A., (1993) Visually Controlled Graphics, *IEEE Trans. Pattern Analysis and Machine Vision*, special issue of computer graphics and computer vision, Vol. 15, No. 6, pp. 602-604

[2] Azarbayejani, A., Horowitz, B., and Pentland, A., (1993) Recursive Estimation of Structure and Motion using the Relative Orientation Constraint, *IEEE Conference on Vision and Pattern Recognition*, NY, NY, June 1993. Also available as: M.I.T. Perceptual Computing Technical Report No. 243.

[3] Azarbayejani, A. and Pentland, A., (1994) Recursive Estimation of Motion, Structure, and Focal Length, *IEEE Trans. Pattern Analysis and Machine Vision*, (to appear).

[4] Darrell, T., and Pentland, A., Space-Time Gestures, (1993) Proc. IEEE Conf. on Computer Vision and Pattern Recognition, NY NY, June 1993.

[5] Darrell, T., Maes, P., Blumberg, B., and Pentland, A., (1994) "A Novel Environment for Situated Vision and Behavior," *IEEE Workshop on Visual Behaviors* pp. 68-72, Seattle, WA., June 19, 1994.

[6] Essa, I., Sclaroff, S., and Pentland, A., (1992) A Unified Approach for Physical and Geometric

Modeling *Computer Graphics Forum*,Vol 2, No. 3, pp. 129-138. *also appears: Eurographics*, Cambridge, England, Sept 7-11.

[7] Essa, I., Darrell, T., and Pentland, A., Modeling and Interactive Animation of Facial Expressions using Vision, *To Appear*: IEEE Conf. Computer Vision and Pattern Recognition, Seattle, WA, June 1994. Also available as: M.I.T. Perceptual Computing Technical Report No. 256

[8] Maes, P., Darrell, T., Blumberg, B., and Pentland, A., The ALIVE System: Full-body Interaction with Animated Autonomous Agents, *appeared: SIGGRAPH-93 Tommorrow's Realities Track*, M.I.T. Media Laboratory Perceptual Computing Technical Report No. 257, January 1994.

[9] Pentland, A., and Williams, J. (1989) Good Vibrations: Modal Dynamics for Graphics and Animation, *ACM Computer Graphics* Vol. 23, No. 4, pp. 215-222, August, 1989.

[10] Pentland, A. and Sclaroff, S., (1991) Closed-Form Solutions For Physically Based Shape Modeling and Recognition *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 13, No. 7, pp. 715-730.

[11] Pentland, A., and Horowitz, B., (1991) Recovery of Non-Rigid Motion and Structure *IEEE Trans. Pattern Analysis and Machine Intelligence* Vol. 13, No. 7, pp. 730-742.

[12] Pentland, A., (1993) Modal Descriptions for Recognition and Tracking, *IEICE Trans. Information and Systems*, Vol. j76-D-II, No. 8, 1489-1496.

[13] Sclaroff, S., and Pentland, A., (1991) Generalized Implicit Functions for Computer Graphics, *ACM Computer Graphics*, Vol. 25, No. 2, pp. 247-250.

[14] Sclaroff, S., Essa, I., and Pentland, A., (1992) Vision-Based Animation: Appications of a Unified Approach for Physical and Geometric Modeling, *Eurographics Workshop on Physically-Based Modeling*, Cambridge, England, Aug 31.

[15] Sclaroff, S., and Pentland, A., (1993) A Modal Framework for Correspondence and Recognition, *Int'l Conference on Computer Vision*, Berlin, Germany, May 1993. Also available as: M.I.T. Perceptual Computing Technical Report No. 201.

[16] Starner, T., and Pentland, A., (1994) Visual Interpretation of Americal Sign Language, M.I.T. Media Laboratory Perceptual Computing Technical Report Number 306.