

Vision for a Smart Kiosk

James M. Rehg Maria Loughlin Keith Waters

Digital Equipment Corporation[†]
Cambridge Research Lab
One Kendall Square, Bldg. 700
Cambridge, MA 02139, USA
rehg,loughlin,waters@crl.dec.com

Abstract

We describe a novel computer vision application: vision-based human sensing for a Smart Kiosk interface. A Smart Kiosk is a free-standing information dispensing computer appliance capable of engaging in public interactions with multiple people. Vision sensing is a critical component of the kiosk interface, where it is used to determine the context for the interaction. We present a taxonomy of vision problems for a kiosk interface and describe a prototype kiosk which uses color stereo tracking and graphical output to interact with several users.

1. Introduction

An automated, information-dispensing *Smart Kiosk*, which is situated in a public space for use by a general audience, poses a challenging human-computer interface problem. We are exploring a social interface paradigm for a Smart Kiosk, in which a graphical, speaking agent is used to output information and communicate cues such as focus of attention. Human sensing techniques from computer vision play a critical role in the kiosk interface. Inobtrusive video cameras provide a wealth of information about users, ranging from their three dimensional location to their body language and facial expressions.

Although vision-based human sensing has received increasing attention in the past five years (see the proceedings [14, 1, 5]) relatively little work has been done on identifying the perceptual requirements for functioning vision-enabled user-interfaces. A few notable exceptions are the pioneering work of Krueger [9], the Mandala Group [8], the Alive system [10], and a small body of work on gesture-based control for desktop and set-top box environments (see [7] and the survey article [13].)

[†]The following are trademarks of Digital Equipment Corporation: Alpha, DEC, DECtalk, and the DIGITAL logo.

In contrast to these earlier vision-enabled systems, the kiosk interface supports *public interaction with multiple users*. It must be able to actively initiate and terminate interactions and manage an uncertain, dynamically-changing user population. Computer vision techniques play a key role in sensing the context for the interaction, which differs substantially from current desktop or immersive VR interfaces. As a result, kiosk interfaces represent a novel application domain for computer vision research.

In this article we characterize the vision task requirements for a kiosk interface, and present experimental results from a working prototype. First we describe the kiosk interface problem and present a taxonomy of visual sensing requirements. Then we present modules for real-time visual sensing (including motion detection, colored object tracking, and stereo ranging) in our prototype kiosk. More details on the kiosk architecture and modules can be found in [24]. Finally, we present an empirical evaluation of the accuracy of our vision algorithms, and describe three interaction experiments with the kiosk prototype.

2. Vision Tasks for a Kiosk Interface

The dynamic, unconstrained nature of a public space, such as a shopping mall, poses three challenges for a kiosk user-interface. First, the kiosk should support a social interaction paradigm, in which communication occurs through speaking, gestures, and graphical output. It is natural that a device in a public place should follow rules for communication and behavior that its users are already familiar with. Kiosk users will come from a broad range of backgrounds and will not have an opportunity for extensive training with the interface.

The second challenge is that the kiosk interface should actively participate in initiating and regulating interactions with its users. It should greet someone who approaches it and engage people at a distance. The third challenge is the

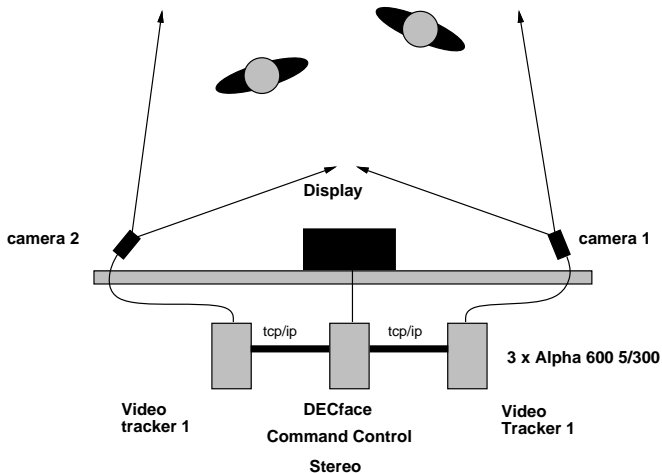


Figure 1. Smart Kiosk prototype. The interaction space lies on one side of a partition, and three DIGITAL Alpha workstations lie on the other. Interactions take place around eight feet from the kiosk display.

ability to interact simultaneously with multiple users, and dynamically adjust to changes in their number. For example, a kiosk dispensing information about local events in a museum should communicate to all interested parties and base the time it spends with a specific user on the number of others who are waiting.

These challenges require sophisticated visual sensing to monitor the kiosk's users and determine the *context* for the interaction. In contrast to the kiosk domain, the designer of a conventional graphical user-interface has a great deal of control over context. Mapping a mouse event to a specific action, for example, is a simple and unambiguous task. A kiosk, on the other hand, must infer the context for interaction by sensing its user's actions. While it is easy to draw up a litany of challenging vision problems to support these goals, we will demonstrate that simple sensing tasks can already enable a variety of interesting interactions.

There are three basic vision tasks that a Smart Kiosk should be capable of:

Detect the arrival and departure of potential users in the vicinity of the kiosk.

Track the position and velocity of users over time.

Identify individual users during a multi-person interaction.

These tasks take place inside the kiosk's *interaction space*, which is illustrated in Figure 1.

Detecting the arrival of a potential user inside the field of view of the kiosk's cameras is a basic vision task. Face

and motion cues seem to be the most promising means for detecting people in image sequences. Our prototype kiosk uses motion, and we plan to add face detection. Once a person has been detected, their position relative to the kiosk must be determined so further actions can be taken.

We employ stereo triangulation on detected image positions in two calibrated cameras to locate the user in the scene. Stereo has the advantage of giving absolute position measurements while placing minimal constraints on the kiosk environment. Another approach is to exploit looming motion or region size in a single camera image to get a crude measure of distance. Alternatively, if the position of the ground plane is known and the user's feet can be detected, then the intersection gives the distance from the camera [26]. Detection and localization are critical in enabling the kiosk interface to respond directly to the approach of a potential user.

Once a person has entered the kiosk's interaction space, a vision-based tracking process is invoked. Tracking is a basic tool in deciding whether a given person is a good candidate for interaction. Individuals whose attention is focused on a specific activity unrelated to the interface may not welcome a show of interest from the kiosk. In general, determining the context for a potential interaction is a complex task which may involve estimating the individual's direction of gaze, facial expression, and body language. We have focused on the most basic cue, the user's velocity relative to the kiosk. A person who is moving slowly towards the kiosk is a more likely candidate for interaction than someone moving quickly on a divergent heading. We employ stereo tracking of color and motion blobs to estimate the position and velocity of an individual. The tracked position of the user can also support further analysis by more complex vision modules.

Multiple users inside the kiosk's interaction space pose additional sensing challenges. The most basic requirement is the ability to tell individual users apart in order to determine when a new user has entered the scene or an existing user has departed. This is a kind of short-term recognition problem for which both clothing color and facial images may be employed. We currently use a color model of each user's shirt to identify them during tracking. Detecting departures can be particularly difficult in the presence of occlusions from multiple users. A second requirement for multi-user kiosks is the ability to communicate focus of attention. For example, our kiosk uses a synthetic talking head to communicate with its audience. By controlling the gaze and pose of the head, we can identify the intended recipients of a given remark. This requires the ability to locate individual users in the scene.

The set of tasks described above provides a core competency for vision-enabled kiosks. In examining the broader spectrum of vision-based sensing tasks, it is useful to con-

sider a taxonomy of problems based on distance from the kiosk. Since the image resolution available for human sensing varies inversely with the square of the distance from the interface, the characteristics of the camera and display place a strong constraint on feasible approaches.

The taxonomy in Table 1 divides distance from the kiosk into three categories. We define *distant* (**Dist**) as greater than 15 feet, *midrange* (**Mid**) as 5 to 15 feet, and *proximate* (**Prox**) as less than 5 feet. The categories are nested, since a task that can be performed at a given distance can usually be performed at a closer one as well. Certain body features can be measured within each category, making it possible to estimate attributes of the figure and enabling certain kiosk behaviors. At far distances, for example, the figure is essentially a location which the kiosk can monitor. As the user approaches, they can be enticed to enter the proximate range, permitting communication which is supported by detailed analysis of face and hand gestures.

	Features	Attributes	Behaviors
Dist	Whole body	Position	Monitor
Mid	Head and torso	Orientation	Entice
Prox	Face/Hands	Expression	Communicate

Table 1. Taxonomy of vision tasks based on increasing proximity to the kiosk.

3. A Prototype Smart Kiosk

In order to explore the interface issues described above, we have developed a prototype Smart Kiosk based on simple vision sensing, a speaking synthetic agent called *DECface*, and some simple behaviors. The kiosk architecture is illustrated in Figure 1. Two cameras cover the kiosk workspace, the half-plane in front of the display, and provide stereo measurements of the user’s position. This section describes the software modules in greater detail.

3.1. Tracking Using Motion and Color Stereo

Color and motion stereo tracking of users provides position and velocity measurements for kiosk interface. We represent each user as a blob in the image plane, and triangulate on the blob centroids from two cameras to localize each user in the environment (see [2] for a related approach.) We use motion stereo for range estimation at far distances, and color stereo for short range tracking of multiple users.

We use a modified version of the color histogram indexing and backprojection algorithm of Swain and Ballard [20] to track multiple people in real-time within approximately



Figure 2. Sample output from the color tracking module. Images were obtained from the right camera of the stereo pair. The left hand portion of the display shows a plan view of the scene with a cross marking the 3D location of the individual projected onto the ground plane.

15 feet of the kiosk. We obtain histogram models of each user through a manual segmentation stage. Like other researchers [26, 11, 8, 20], we have found normalized color to be a descriptive, inexpensive, and reasonably stable feature for human tracking. We use local search during tracking to improve robustness and speed. To avoid color matches with static objects in the environment, background subtraction is used to identify moving regions in the image before indexing. Sample output from the color tracker is illustrated in Figure 2.

Given motion- or color-based tracking of a person in a pair of calibrated cameras, stereo triangulation is used to estimate the user’s 3D location. In our kiosk prototype, we use a pair of verged cameras with a six foot baseline. Extrinsic and intrinsic camera parameters are calibrated using a non-linear least-squares algorithm [21]. Given blob centroids in two images, triangulation proceeds through ray intersection. The 3D position is chosen as the point of closest approach in the scene to the rays from the two cameras that pass through the detected centroid positions. We use a simple Kalman filter to estimate the position and velocity of the user in the plane of the floor.

3.2. DECface Agent

DECface, a talking synthetic face agent [23], is the primary feedback mechanism for the kiosk. It combines real-time computer-generated face imagery with the audio output of a speech synthesizer. Figure 3 shows some sample

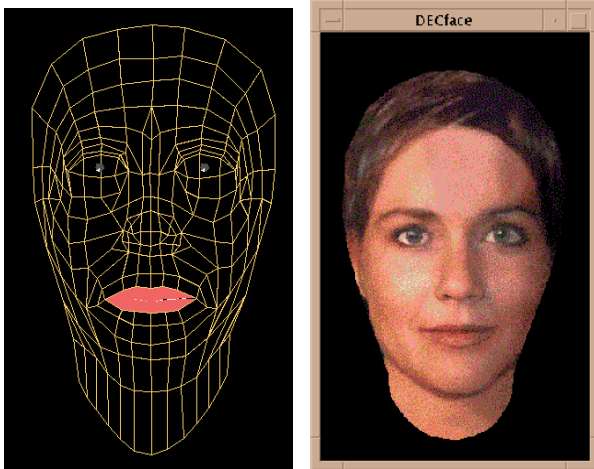


Figure 3. DECface rendered in wireframe (left) and as a texture mapped female face (right).

DECface output. DECface has three capabilities that are ideally suited to constructing a reactive agent: the ability to speak an arbitrary piece of text at a specific speech rate in one of eight voices from one of eight faces, the creation of simple facial expressions under control of a facial muscle model [22], and simple head and eye rotation. We employ the DECTalk [6] speech synthesizer in our experiments.

The behaviors of our prototype kiosk are constructed from three simple DECface output mechanisms: it can speak text to its audience, it can control its head orientation and eye gaze, it can blink its eyes. Head and eye motion are used to communicate focus of attention, while eye blinks are an example of a reflexive behavior. A finite state machine specifies the connection between the vision module outputs and DECface's output. We currently construct the behavior modules by hand for each interaction scenario. More details on the implementation of kiosk behavior can be found in [24].

3.3. Implementation

The kiosk prototype is implemented as a set of independent software modules (threads) running on a network of workstations and communicating by message-passing over TCP/IP sockets. We currently have five types of modules: motion blob detection, color tracking, stereo triangulation, DECface, and behavior. Figure 1 illustrates the hardware configuration used in the kiosk prototype. All of the experiments in this paper used three DIGITAL Alpha workstations. Two of the workstations were used for the two color or blob tracking modules, and the third was used for the DECface, stereo, behavior, and routing modules. Images were acquired from two Sony DXC-107 color CCD cam-

eras and digitized with two DIGITAL Full Video Supreme digitizers.

The network architecture supports both direct socket connections between modules and communication via a central routing module. At initialization, all modules connect to the router, which maps module names to IP addresses and can log message transmissions for debugging purposes. The router limits the complexity of the network connections and supports on-the-fly addition and removal of modules. In cases where maximum network throughput is important, as when the output of color stereo tracking is driving DECface gaze behavior, a direct connection between modules is established.

4. Experimental Results

We conducted two sets of experiments with our prototype Smart Kiosk. The first set was designed to measure the accuracy of the color-based stereo position and velocity estimation. The second set consisted of three real-time interactions with the kiosk prototype by both one and two users.

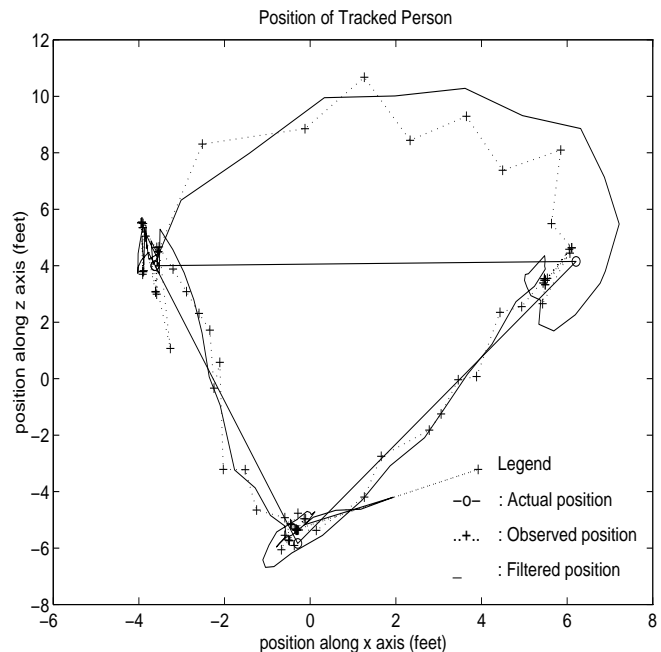


Figure 4. Triangular path of a single user, showing the raw and Kalman filtered stereo output superimposed with the ground truth.

4.1. Accuracy Experiments

We conducted two experiments with ground truth data to determine the effectiveness of the color stereo tracking system. In the first, a single user walked along a triangular path on the floor while being tracked by the color stereo system. The estimated and ground truth trajectories are given in Figure 4. We filtering the estimated x, y, z coordinates with a constant acceleration Kalman filter (Sec. 2.7 in [3]) to improve the smoothness of the trajectory. These results demonstrate good qualitative agreement with the actual motion, particularly for distances within fourteen feet of the kiosk.

The larger errors along the arc from A to B are due to the decreased size of the target at this distance, and darker lighting conditions in that part of the room. These effects cause the blob mass following histogram intersection to become scattered, and reduce the quality of the centroid estimate. In general we found the centroid localization to be quite noisy and sensitive to changes in illumination.

Vertices:	A	B	C
X	1.59	7.50	10.07
Y	3.55	3.96	6.70
Z	9.70	14.80	15.90

Table 2. Standard error deviations in inches of estimated user positions for vertices A, B, and C in the triangular path from Figure 4.

The user also paused for several frames at each node of the triangle, making it possible to estimate the error variance in the measured position. Table 2 shows the standard deviations in inches in the X, Y, and Z coordinates for the triangulated blob centroid. In this case, the X and Z axes lie in the plane of the floor, with the positive Z axis measuring increasing distance from the kiosk, while the Y axis points into the floor. These deviations vary with position on the floor, as might be expected due to changing illumination and viewing angle. Not surprisingly, variation was largest in the depth direction (Z axis) where the stereo constraint is the weakest.

In the second experiment, a user approached the kiosk on a straight line passing through the display. At a point about 10 feet from the kiosk they abruptly veered off, and passed the kiosk on their right. We extracted estimates of the magnitude and angle of the user’s velocity by Kalman filtering the raw x, z measurements. The resulting trajectories are plotted in Figure 5. The change in angle around time 2.5 indicates reasonable sensitivity to the user’s motion. A simple threshold on angle in this case would be sufficient to

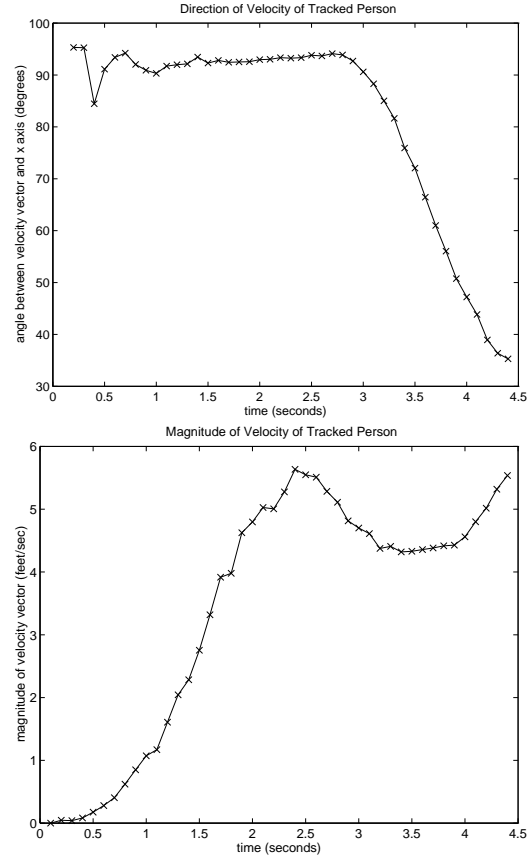


Figure 5. Heading angle and magnitude of the users velocity during a two stage approach to the kiosk. From time 0 to 2.5, the user is on a bearing aimed directly at the kiosk. At time 2.5, the user veers onto a tangential course for the remainder of the series.

discriminate users who are heading towards the kiosk from those who are not.

4.2. Interaction Experiments

We conducted three experiments in real-time vision-directed behavior on our prototype kiosk. The first experiment used proximity sensing in conjunction with some simple behavioral triggers to detect a single, distant user and entice him or her to approach the kiosk. The user was detected independently in two cameras using image differencing, and motion stereo provided estimates of the person’s distance from the kiosk. This information was sent to the behavioral module. The range of 3D detection was fairly large, beginning at approximately seventy feet and ending a few feet away from the kiosk. For this experiment we implemented a simple trigger behavior which divided the



Figure 6. Three frames of a view through a Handicam while DECface tracks a user in 3D using color stereo.

workspace into near, middle, and far regions, and associated a set of sentences to the transitions between the regions. As the user approached the kiosk, the behavior model detected the transitions between regions and caused DECface to speak an appropriate message.

The second experiment explored the use of close range tracking to drive DECface gaze behavior. A single user was tracked using the color stereo algorithm described earlier. The user's 3D position was converted into a gaze angle in DECface's coordinate system and used to control the x -axis orientation of the synthetic face display in real-time. We implemented a simple gaze behavior which enabled DECface to follow the user with its gaze as the user roamed about the workspace. Figure 6 shows three frames of the display from the user's viewpoint as he walks past the kiosk from left to right.

The third experiment built upon the vision-directed gaze behavior to communicate the kiosk's focus of attention when interacting with multiple users. For this example we implemented a very simple "storytelling" behavior for an audience of two persons. A six sentence monologue is delivered by DECface to one user, and is interspersed with side comments that are directed at the second user. We used the direction of DECface's gaze to indicate the recipient of each sentence, and employed 3D color stereo tracking to update the gaze direction in real-time as the users changed position. Figure 7 shows a snapshot of the audience during the story-telling experiment.

5. Previous Work

There are two bodies of work which are closely related to our Smart Kiosk system. The first is investigations into vision-based user-interfaces for desktop computing [16], set-top boxes [7], and virtual environments [9, 8, 19, 11, 10, 13]. In particular, the Alive system [10], and the works that preceded it [9, 8], employed vision sensing to allow users to interact with autonomous agents in a virtual environment. In contrast to the virtual environment scenario, the kiosk interface is embedded in the public environment of its users, and must conform to norms about acceptable behav-



Figure 7. 3D color tracking of two individuals during the "storytelling" sequence. During the sequence the two individuals exchange locations.

ior. As a result, the problem of determining the context for the interaction is the most critical sensing task required by the interface. An preliminary report on this work appears in [25].

The second body of related work concerns algorithms for tracking human motion using video images [15, 12, 17, 4, 18, 26, 2]. Our color and motion blob algorithms are most closely related to those of Wren *et al.* [26], which are employed in the Alive system. The color histogram representation for blobs [20] that we employ is more descriptive than a single color blob model and therefore more appropriate to our task of discriminating between multiple users. Related work on blob stereo tracking is described in [2].

6. Conclusions and Future Work

We have demonstrated a significant role for visual sensing in public user-interfaces. We have described real-time multi-user experiments with a prototype Smart Kiosk, and have investigated the error behavior of our tracking algorithm. We have found, as others have, that color is a valuable feature for tracking people in real-time. We have shown that it can be used in conjunction with stereo to measure a user's 3D position and velocity.

The key to an effective kiosk interface is natural communication with its users within the context of their environment. We plan to extend our initial results through improved sensing technology and more complex and compelling kiosk behaviors. Our visual sensing work has focused on detecting and tracking people in the distance and

at midrange, to support the initiation and control of interactions. We plan to develop close-range sensing to identify users' facial expressions and gaze. We are also interested in adding other sensing modalities such as speech.

The second focus of our future work is the development of compelling kiosk behaviors. We can develop behavioral characteristics for DECface's voice, speech pattern, facial gestures, head movement and expressions that will cause users to attribute a personality to the kiosk. We would also like the kiosk to create goals dynamically, based on its charter, user input, and the direction of the current interaction. These goals drive the motivational actions of the kiosk. Management of competing goals and flexibility in response to a changing user population are key technical challenges.

References

- [1] J. Aggarwal and T. Huang, editors. *Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, TX, November 1994. IEEE Computer Society Press.
- [2] A. Azarbayejani and A. Pentland. Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features. Technical Report 363, MIT Media Lab, Perceptual Computing Section, January 1996.
- [3] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.
- [4] A. Baumberg and D. Hogg. An efficient method for contour tracking using active shape models. In J. Aggarwal and T. Huang, editors, *Proc. of Workshop on Motion of Non-Rigid and Articulated Objects*, pages 194–199, Austin, Texas, 1994. IEEE Computer Society Press.
- [5] M. Bichsel, editor. *Intl. Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, June 1995.
- [6] Digital Equipment Corporation. *DECTalk Programmers Reference Manual*, 1985.
- [7] W. Freeman and C. Weissman. Television control by hand gestures. In M. Bichsel, editor, *Proc. of Intl. Workshop on Automatic Face and Gesture Recognition*, pages 179–183, Zurich, Switzerland, June 1995.
- [8] M. Group. Mandala: Virtual village. In *SIGGRAPH-93 Visual Proceedings*, 1993.
- [9] M. Krueger. *Artificial Reality II*. Addison Wesley, 1990.
- [10] P. Maes, T. Darrell, B. Blumberg, and A. Pentland. The ALIVE system: Wireless, full-body interaction with autonomous agents. *ACM Multimedia Systems*, Spring 1996. Accepted for publication.
- [11] C. Maggioni. Gesturecomputer – New ways of operating a computer. In *Proc. of Intl. Workshop on Automatic Face and Gesture Recognition*, pages 166–171, Zurich, Switzerland, June 1995.
- [12] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):580–591, 1993.
- [13] V. Pavlović, R. Sharma, and T. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. Technical Report UIUC-BI-AI-RCV-95-10, University of Illinois at Urbana-Champaign, December 1995.
- [14] A. Pentland, editor. *Looking at People Workshop*, Chambéry, France, August 1993. IJCAI.
- [15] A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):730–742, 1991.
- [16] J. M. Rehg and T. Kanade. Digiteyes: Vision-based hand tracking for human-computer interaction. In J. K. Aggarwal and T. S. Huang, editors, *Proc. of Workshop on Motion of Non-Rigid and Articulated Objects*, pages 16–22, Austin, Texas, 1994. IEEE Computer Society Press.
- [17] J. M. Rehg and T. Kanade. Visual tracking of high dof articulated structures: an application to human hand tracking. In J.-O. Eklundh, editor, *Proceedings of European Conference on Computer Vision*, volume 2, pages 35–46, Stockholm, Sweden, 1994. Springer-Verlag.
- [18] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. of Fifth Intl. Conf. on Computer Vision*, pages 612–617, Boston, MA, 1995. IEEE Computer Society Press.
- [19] J. Segen. Gest: A learning computer vision system that recognizes hand gestures. In R. Michalski and G. Tecuci, editors, *Machine Learning IV*. Morgan Kaufmann, 1993.
- [20] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.
- [21] R. Szeliski and S. B. Kang. Recovering 3D shape and motion from image streams using nonlinear least squares. *Journal of Visual Communication and Image Representation*, 5(1):10–28, March 1994.
- [22] K. Waters. A muscle model for animating three-dimensional facial expressions. *Computer Graphics (SIGGRAPH '87)*, 21(4):17–24, July 1987.
- [23] K. Waters and T. Levergood. An automatic lip-synchronization algorithm for synthetic faces. *Multimedia Tools and Applications*, 1(4):349–366, Nov 1995.
- [24] K. Waters, J. M. Rehg, M. Loughlin, S. B. Kang, and D. Terzopoulos. Visual sensing of humans for active public interfaces. Technical Report CRL 96/5, Digital Equipment Corp. Cambridge Research Lab, 1996.
- [25] K. Waters, J. M. Rehg, M. Loughlin, S. B. Kang, and D. Terzopoulos. Visual human sensing for active public interfaces. In S. Pentland and R. Cipolla, editors, *Computer Vision in Man-Machine Interfaces*. Cambridge University Press, Cambridge, UK, 1997.
- [26] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. Technical Report 353, MIT Media Lab, Perceptual Computing Section, 1995.