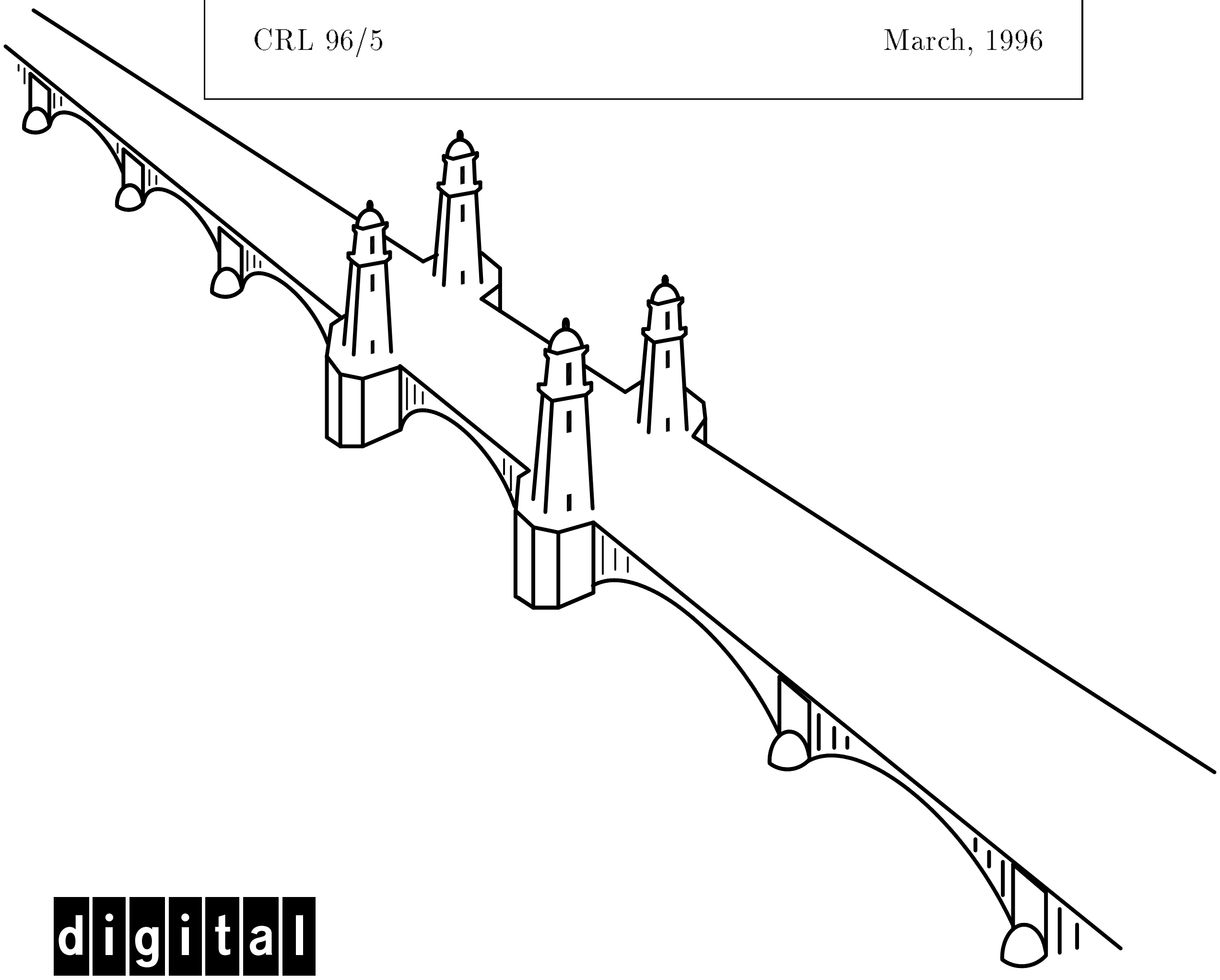
Visual Sensing of Humans for Active Public Interfaces

K. Waters, J. Rehg, M. Loughlin, S. B. Kang,
and D. Terzopoulos

Digital Equipment Corporation
Cambridge Research Lab

CRL 96/5 March, 1996



CAMBRIDGE RESEARCH LABORATORY
Technical Report Series

Digital Equipment Corporation has four research facilities: the Network Systems Laboratory, the Systems Research Center, and the Western Research Laboratory, all in Palo Alto, California; and the Cambridge Research Laboratory, in Cambridge, Massachusetts.

The Cambridge laboratory became operational in 1988 and is located at One Kendall Square, near MIT. CRL engages in computing research to extend the state of the computing art in areas likely to be important to Digital and its customers in future years. CRL's main focus is applications technology; that is, the creation of knowledge and tools useful for the preparation of important classes of applications.

CRL Technical Reports can be ordered by electronic mail. To receive instructions, send a message to one of the following addresses, with the word **help** in the Subject line:

On Digital's EASYnet:	CRL::TECHREPORTS
On the Internet:	techreports@crl.dec.com

digital™

Cambridge Research Laboratory
One Kendall Square
Cambridge, Massachusetts 02139

Visual Sensing of Humans for Active Public Interfaces

K. Waters, J. Rehg, M. Loughlin, S. B. Kang, and D. Terzopoulos

Digital Equipment Corporation
Cambridge Research Lab

CRL 96/5 March, 1996

Abstract

Computer vision-based sensing of people enables a new class of *public* multi-user computer interfaces. Computing resources in public spaces, such as automated, information-dispensing kiosks, represent a computing paradigm that differs from the conventional desktop environment and, correspondingly, require a user-interface metaphor quite unlike the traditional WIMP interface. This paper describes a prototype public computer interface, which employs color and stereo tracking to sense the users' activity and an animated, speaking agent to attract attention and communicate through visual and audio modalities.

URL: *http://www.research.digital.com/CRL/projects/vision-graphics*

Keywords: Agents, Blob Tracking, Color Tracking, Facial Animation, Human-Computer Interaction, Smart Kiosk, Stereo

Contents

List of Figures

List of Tables

1 Introduction

An automated, information-dispensing Smart Kiosk, which is situated in a public space for use by a general clientele, poses a challenging human-computer interface problem. A *public* kiosk interface must be able to actively initiate and terminate interactions with users and divide its resources among multiple customers in an equitable manner. This interaction scenario represents a significant departure from the standard WIMP (windows, icons, mouse, pointer) paradigm, but will become increasingly important as computing resources migrate off the desktop and into public spaces. We are exploring a social interface paradigm for a Smart Kiosk, in which computer vision techniques are used to sense people and a graphical, speaking agent is used to output information and communicate cues such as focus of attention.

Human sensing techniques from computer vision can play a significant role in *public user-interfaces* for kiosk-like appliances. Using unobtrusive video cameras, they can provide a wealth of information about users, ranging from their three dimensional location to their facial expressions and body language. Although vision-based human sensing has received increasing attention in the past five years (see the proceedings [18, 1, 5]) relatively little work has been done on integrating this technology into functioning user-interfaces. A few notable exceptions are the pioneering work of Krueger [13], the Mandala Group [9], the Alive system [14], and a small body of work on gesture-based control for desktop and set-top box environments (see [17] for a survey.)

This chapter describes our prototype kiosk and some experiments with vision-based sensing. The kiosk prototype currently consists of a set of software modules that run on several workstations and communicate through message-passing. It includes modules for real-time visual sensing (including motion detection, colored object tracking, and stereo ranging), a synthetic agent called DECface [27], and behavior-based control.

We describe this architecture and its implementation in the following sections and present experimental results related to proximity-based interactions and active gaze control. Section 2 describes some characteristics of the user-interface problem for public kiosk-like devices. Section 3 discusses and describes how computer vision can be used in a public user-interface. Section 3.2 presents persistence behavior tracking for proximate, midrange and distant interactions as well as stereo tracking using color. Section 3.3 describes the feedback technology of DECface. Section 3.4 describes the behaviors we are interested in developing. Section 4 describes implementation details of our prototype kiosk. Section 5 reports on vision-directed behavior experiments performed with the prototype. Section 6 reviews previous work in vision-based human sensing. Section 7 describes future work and Section 8 concludes the paper.

2 Characteristics of Public User-Interfaces

The dynamic, unconstrained nature of a public space, such as a shopping mall, poses a challenging user-interface problem for a Smart Kiosk. We refer to this as the *public user-interface* problem, to differentiate it from interactions that take place in structured, single-user desktop [23] or virtual reality [14] environments. We have developed a prototype Smart Kiosk which we are using to explore the space of effective public interactions. Our prototype has three functional components: human sensing, behavior, and graphical/audio output. This section outlines some basic requirements for public interactions and their implications for the components of a Smart Kiosk.

The effectiveness of a Smart Kiosk in reaching its target audience can be measured by two quantities, *utilization* and *exposure*. Utilization refers to the percentage of time the kiosk is in use, while exposure refers to the number of different users that interact with the system during some unit of time. Maximizing utilization ensures the greatest return on the cost of kiosk technology, while maximizing exposure prevents a small number of users from monopolizing the kiosk's resources. We are exploring interaction models that are *human-centered*, *active*, and *multi-user*, as a way of maximizing kiosk utilization and exposure.

We are interested in human-centered interactions, in which communication occurs through speaking and gestures. Since a Smart Kiosk is situated in its users' physical environment, it is natural for it to follow rules of communication and behavior that users can interpret immediately and unambiguously. This is particularly important given the broad range of backgrounds of potential users for Smart Kiosks, and the lack of opportunity for user training. Human-centered output can be obtained through the use of a graphical, speaking agent, such as DECface. Human-centered input consists of visual and acoustical sensing of human motion, gesture, and speech. The behavior component plays a critical role in regulating the communication between the agent and the users, and setting the users' expectations for the interaction.

Active participation in initiating and regulating interactions with users is the second vital task for public interfaces. By enticing potential users, a Smart Kiosk can maximize its utilization, for example, by detecting the presence of a loiterer and calling out to them. This requires the ability to visually detect potential customers in the space around the kiosk and evaluate the appropriateness of beginning an interaction. Active interfaces must behave in socially acceptable fashion. A kiosk interface that is too aggressive could alienate rather than attract potential users. Thus an active interface requires both vision techniques for sensing potential users and behavioral models for interactions.

In addition to being active and human-centered, a kiosk situated in a public space must also be able to deal with the conflicting requests of multiple users. Two basic issues arise when the resources of the interface are shared by multiple users. First, the interface must be able to communicate its focus of attention to the user population, so the center of the interaction is known by all users at all times. The eye gaze of the DECface agent provides a natural mechanism for communicating focus of attention. Effective gaze behavior requires a graphical display with compelling dynamics and a model of gaze behavior. Gaze behavior is designed both to indicate the focus of attention and to draw the audience into the interaction through eye contact.

Besides communicating a focus of attention, the interface must build and maintain a representation of its audience to support multi-user interactions. At the minimum, it must track and identify the position of its current users, so it can recognize arrivals and departures. This makes it possible to include new users into the interaction in an efficient manner and ensure a fair allocation of resources. In addition to visual sensing capabilities, a multi-user interface must have a behavioral model of the interaction process that allows it to monitor its allocation of resources to its community of users and make tradeoffs between conflicting demands.

3 The Smart Kiosk Interface

In order to explore the public interaction issues described above, we have developed a prototype Smart Kiosk interface based on visual sensing, a speaking DECface agent, and some simple behaviors. This section describes the components of the interface and their design.

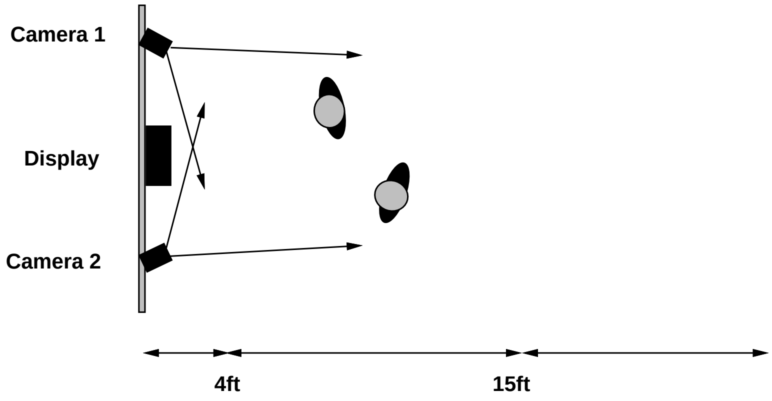


Figure 1: Interaction space for a Smart Kiosk. Two cameras monitor the position of multiple users in front of the kiosk display.

3.1 Designing the Interaction Space

We currently assume that the kiosk display is the focal point of an interaction with multiple users which will take place in a space ranging from the front of the display out to a few tens of feet. Users walking into this space will be imaged by one or more cameras positioned around the kiosk display. Figure 1 illustrates the sensor and display geometry for our kiosk prototype.

We can characterize the effect of camera and display positioning on the visual sensing tasks for the kiosk interface. The image resolution available for human sensing varies inversely with the square of the distance from the interface. As a result, it is possible to roughly classify the human behaviors, articulation, and features that can be measured from images into categories based on proximity to the interface. Table 1 defines a taxonomy in which proximity dictates the type of processing that can take place. These categories are nested, since a task that can be done at some distance can always be performed at a closer one. In our current system we define "proximate" as less than about 4 feet, "midrange" as from 4 to 15 feet, and "distant" as greater than 15 feet.

3.2 Person Tracking Using Motion and Color Stereo

Color and motion stereo tracking provide visual sensing for the Smart Kiosk prototype. We represent each user as a blob in the image plane, and trian-

	Proximate	**Midrange**	**Distant**
Human features	Face/Hands	Head and torso	Whole body
Human articulation	Expression	Orientation/Pose	Walking
Human behaviors	Emotion	Attention	Locomotion

Table 1: Taxonomy of visual sensing for public user interfaces based on distance from the focal point of interaction.

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

gulate on the blob centroids from two cameras to localize each user in the environment (see [3] for a related approach.) Stereo tracking provides 3D localization of users relative to the kiosk display. This information can be used to initiate interactions based on distance from the kiosk and to provide DECface with cues for gaze behavior in a multi-user setting.

We use motion-based blob detection to localize a single person at a long range from the kiosk. We assume that the user is the only moving object in the scene and employ standard image differencing to identify candidate blob pixels. A bounding box for blob pixels is computed by processing the difference image one scanline at a time and recording the first "on" pixel and the last "off" pixel value above a predefined threshold (see Figure 2 for an example). This simple approach is very fast, and has proven useful in our current kiosk environment.

We use a modified version of the color histogram indexing and back-projection algorithm of Swain and Ballard [24] to track multiple people in real-time within approximately 15 feet of the kiosk. We obtain histogram models of each user through a manual segmentation stage. Like other researchers [28, 15, 9, 24], 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. We use motion stereo for proximity sensing at far distances, and color stereo for short range tracking of multiple users. 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 [25] and a planar target [12]. 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.

3.3 Feedback: DECface

DECface, a talking synthetic face, is the visual complement of the speech synthesizer DECtalk [7]. Where DECtalk provides synthesized speech, DECface provides a synthetic face [27]. By combining the audio functionality of a speech synthesizer with the graphical functionality of a computer-generated face, it is possible to create a *real-time* agent as illustrated in Figure 3.

DECface has been created with the following key attributes for our reactive agent: an 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 [26], and simple head and eye rotation.

3.4 Behavior

The behavior module dictates the actions that the smart kiosk carries out in response to internal and external events. The behavioral repertoire of

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

the kiosk is coded as a collection of behavior routines. Individual behavior routines are executed by an action selection arbitrator which decides what to do next given the internal state of the kiosk and the external stimuli that it receives.

The behavior routines generally exercise control over the sensing and DECface modules. A behavior routine will, in general, direct the vision module to acquire perceptual information that is relevant to the particular behavior. It will also direct the DECface module to produce an audiovisual display to the outside world in accordance with the behavior. The behavior routines are organized in a loose hierarchy with the more complex behavior routines invoking one or more primitive routines.

It is useful to distinguish two types of behavior—reflexive behavior and motivational behavior. Reflexive behaviors are predetermined responses to internal conditions or external stimuli. A simple reflexive behavior is the awakening behavior which is triggered when a dormant kiosk senses movement in its territory. Another example is the eye blinking behavior. Eye blinks are triggered periodically so that DECface's eyes exhibit some natural liveliness. A somewhat more complex reflexive behavior determines the detailed actions of the eyes and head when the gaze is redirected. Psychophysical studies of the human oculomotor system reveal that eye and head motions are coupled, with the relatively larger mass of the head resulting in longer transients compared to those of the eyes [6].

By contrast, a motivational behavior is determined by the internal "mental state" of the kiosk, which will in general encode the emotional condition of the kiosk and any task directed plans that it may have. For example, when communicating with a user, DECface is motivated to look at the person. Thus gaze punctuates the interaction [2]. This gaze behavior combines sensed information about the user's current location with predefined rules about the role of gaze in human interactions. As a more elaborate example, the kiosk may be programmed with a good sense of humor and this would motivate it to attract a group of people and tell them jokes. The joke behavior routine would call upon more primitive behavior routines, including gaze control, to talk to different people in the group and keep everyone engaged in the discussion.

An effective strategy for implementing a behavioral repertoire is to first implement a substrate of simple reflexive behaviors before proceeding with the implementation of increasingly complex motivational behaviors. The behavioral repertoire of the kiosk determines its personality in public interaction. For example, smiling, receptive behaviors give the impression of a friendly kiosk. Alternatively, abrupt, challenging behaviors create the impression of a hostile kiosk.

4 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 4 illustrates the hardware configuration used in the kiosk prototype. All of the experiments in this paper used three Digital Alpha[1] 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 cameras 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.

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

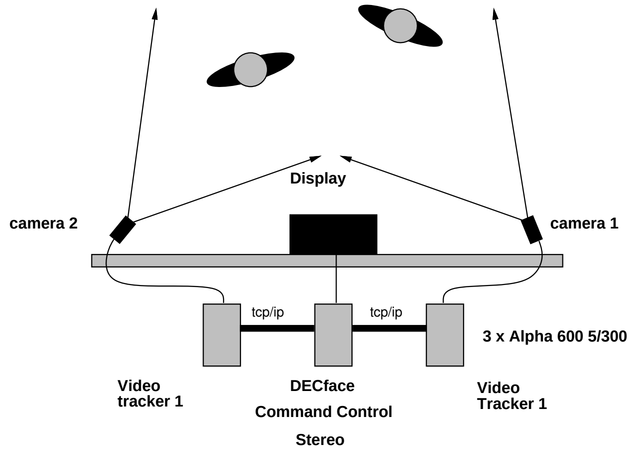


Figure 4: Smart Kiosk prototype. A 24 bit color display is positioned on one side of a partition and three Digital Alpha workstations on the other.

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.

5 Experimental Results

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 the real-time motion blob algorithm described earlier. Stereo triangulation on the blob centroids 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 divides the workspace into near, middle, and far regions, and associates a set of sentences to the transitions between regions. As the user's distance from the kiosk changed, the behavior model detected the transitions between regions and caused DECface

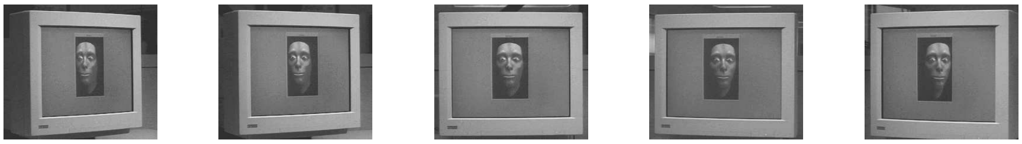


Figure 5: Five frames of a view through a Handicam while DECface tracks a user in 3D using color.

Figure 6: 3D color tracking of two individuals during the "storytelling" sequence. During the sequence the two individuals exchange locations.

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 5 shows five 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 above to show the kiosk's focus of attention when communicating 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 change positions. Figure 6 shows two snapshots of the audience during the story-telling experiment.

6 Previous Work

There are two bodies of work that relate closely to the Smart Kiosk system. The first are investigations into vision-based interfaces for desktop computing [20], set-top boxes [8], and virtual environments [13, 9, 23, 15, 14, 17]. In particular, the Alive system [14], and the works that preceded it [13, 9], have explored the use of vision sensing to support interactions with autonomous agents in a virtual environment.

The second body of related work is on algorithms for tracking human motion using video images [19, 16, 21, 4, 22, 28, 3]. Our color and motion blob algorithms are most closely related to those of Wren *et al.* [28], which are employed in the Alive system. The color histogram representation for blobs [24] that we employ is more descriptive than their single color blob model and therefore more appropriate to our task of identifying multiple users based on color alone. We use stereo for depth recovery rather than the ground plane approach used in Alive because we do not want to segment the entire body or rely on the visibility of the user's feet (also see [3]).

7 Future Work

The key to an effective public interface is natural communication between kiosk and users within the framework of the users' world. There are many ways in which we can develop our kiosk to approach this goal. We will focus on two aspects: (1) improving the users' communication with the kiosk through vision and other modalities, and (2) developing more compelling kiosk behaviors.

Our visual sensing work has been focussed 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. This will allow the kiosk to become more aware of users' intentions and mental state.

Our prototype kiosk senses and tracks people in a simple open environment. A fully developed kiosk may be situated in environments as diverse as airports, shopping malls, theme parks, hotels, and cinema lobbies. In these situations the level of interaction between the user and the kiosk can be enhanced if the kiosk has at its disposal a model of its environment. By determining through stereo the current location of the user relative to itself,

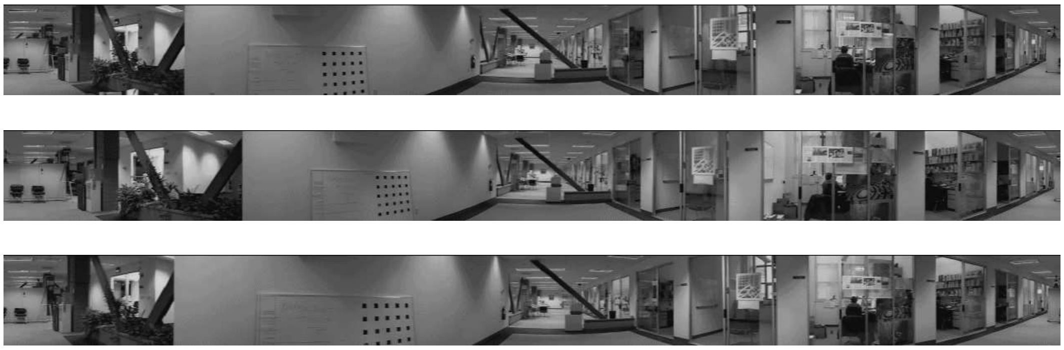


Figure 7: Three panoramic views of the kiosk space scene.

the kiosk can situate the user relative to the model of its environment and respond or react more intelligently to the user's actions.

To this end, we have developed an algorithm to reconstruct the scene using multiple panoramic (full 360° horizontal field of view) images of the scene (Figure 7). The 3D model of the scene is recovered by applying stereo on the multiple panoramic views to create a 3D point distribution (Figure 8(left)) [11]. This 3D point distribution is then used to create a 3D mesh that is texture-mapped with a color panoramic image to produce a 3D reconstructed scene model (Figure 8(right)) [10]. We plan to incorporate models created using this method into the kiosk that we are developing.

We also plan to add alternate input modalities to our kiosk. Speech understanding will enable a user to interact with the kiosk in a direct way. The combination of speech and visual sensing will provide a rich and natural communication medium.

The second focus of our future work is development of more complex and more 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 will be key.

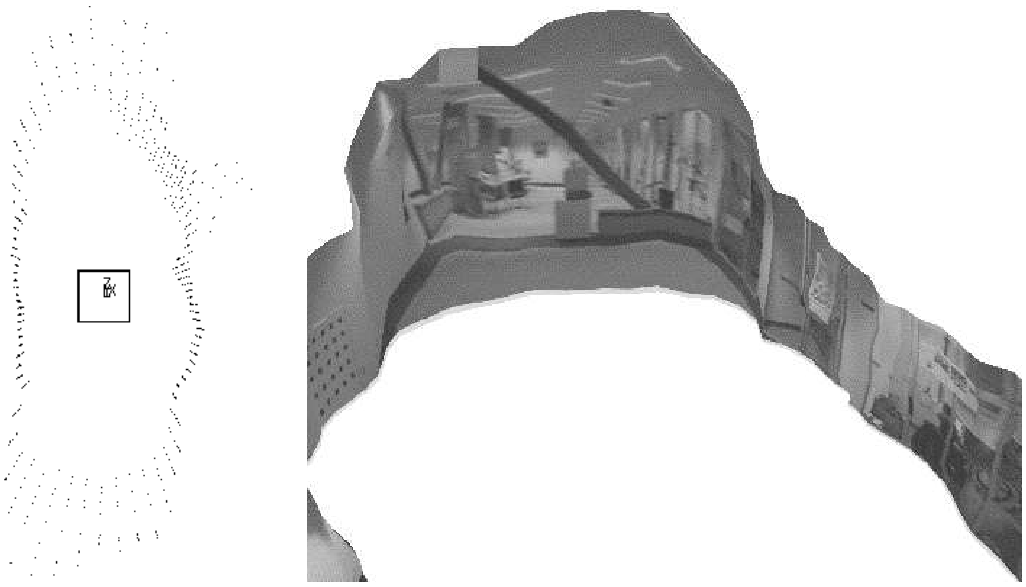


Figure 8: Top view of recovered 3D point distribution (left) and portion of texture mapped reconstructed 3D scene model (right).

8 Conclusion

We have demonstrated a significant role for visual sensing in public user-interfaces. Using simple vision and graphics technology we have developed an engaging user-interface capable of reacting directly to an individual's actions. In addition, we have begun to explore the role of gaze in communicating intention and focus of attention through the use of a synthetic character with an articulate face.

Like other researchers we have found that color is a valuable feature for tracking people in real-time, and that it can be used in conjunction with stereo resolve the users' 3D location.

Acknowledgments

We would like to thank Tamer Rabie of the University of Toronto for making his color-based object tracking software available for our use.

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] M. Argyle and M. Cook. *Gaze and Mutual Gaze.* Cambridge University Press, Cambridge, UK, 1985.

[3] 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.

[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. *Int. Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, June 1995.

[6] R.H.S. Carpenter. *Movements of the Eyes.* Pion Limited, 1972.

[7] Digital Equipment Corporation. *DECtalk Programmers Reference Manual*, 1985.

[8] 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.

[9] Mandala Group. Mandala: Virtual village. In *SIGGRAPH-93 Visual Proceedings*, 1993.

[10] S. B. Kang, A. Johnson, and R. Szeliski. Extraction of concise and realistic 3-D models from real data. Technical Report 95/7, Digital Equipment Corporation, Cambridge Research Lab, October 1995.

[11] S. B. Kang and R. Szeliski. 3-D scene data recovery using omnidirectional multibaseline stereo. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 364–370, June 1996.

[12] S. B. Kang, J. Webb, L. Zitnick, and T. Kanade. A multibaseline stereo system with active illumination and real-time image acquisition. In *Fifth International Conference on Computer Vision (ICCV'95)*, pages 88–93, Cambridge, MA, June 1995.

[13] M. Krueger. *Artificial Reality II.* Addison Wesley, 1990.

[14] 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.

[15] C. Maggioni. Gesturecomputer – New ways of operating a computer. In *Proc. of Intl. Workshop on Automatic Face and Gesture Recognition*, pages 166–171, June 1995.

[16] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):580–591, 1993.

[17] 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.

[18] A. Pentland, editor. *Looking at People Workshop*, Chambery, France, August 1993. IJCAI.

[19] A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(7):730–742, 1991.

[20] J. Rehg and T. Kanade. DigitEyes: Vision-based hand tracking for human-computer interaction. In J. Aggarwal and T. Huang, editors, *Proc. of Workshop on Motion of Non-Rigid and Articulated Objects*, pages 16–22, Austin, TX, 1994. IEEE Computer Society Press.

[21] J. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: An application to human hand tracking. In J. Eklundh, editor, *Proc. of Third European Conf. on Computer Vision*, volume 2, pages 35–46, Stockholm, Sweden, 1994. Springer-Verlag.

[22] J. 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.

[23] J. Segen. Controlling computers with gloveless gestures. In *Proc. Virtual Reality Systems Conf.*, pages 2–6, March 1993.

[24] M. Swain and D. Ballard. Color indexing. *Int. J. Computer Vision*, 7(1):11–32, 1991.

[25] 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.

[26] K. Waters. A muscle model for animating three-dimensional facial expressions. *Computer Graphics (SIGGRAPH '87)*, 21(4):17–24, July 1987.

[27] K. Waters and T. Levergood. An automatic lip-synchronization algorithm for synthetic faces. *Multimedia Tools and Applications*, 1(4):349–366, Nov 1995.

[28] 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.