M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 511
to appear in 1st International Workshop on Managing Interactions in Smart Environments,
December 13-14 1999, Dublin, Ireland

Combining Audio and Video in Perceptive Spaces

Christopher R. Wren, Sumit Basu, Flavia Sparacino, Alex P. Pentland

Perceptual Computing Section, The MIT Media Laboratory ; 20 Ames St., Cambridge, MA 02139 USA
{cwren,flavia,sbasu,sandy}@media.mit.edu
http://www.media.mit.edu/vismod/

Abstract

Virtual environments have great potential in applications such as entertainment, animation by example, design interface, information browsing, and even expressive performance. In this paper we describe an approach to unencumbered, natural interfaces called Perceptive Spaces with a particular focus on efforts to include true multi-modal interface: interfaces that attend to both the speech and gesture of the user. The spaces are unencumbered because they utilize passive sensors that don't require special clothing and large format displays that don't isolate the user from their environment. The spaces are natural because the open environment facilitates active participation. Several applications illustrate the expressive power of this approach, as well as the challenges associated with designing these interfaces.

1 Introduction

We live in real spaces, and our most important experiences are interactions with other people. We are used to moving around rooms, working at desktops, and spatially organizing our environment. We've spent a lifetime learning to competently communicate with other people. Part of this competence undoubtedly involves assumptions about the perceptual abilities of the audience. This is the nature of people.

It follows that a natural and comfortable interface may be designed by taking advantage of these competences and expectations. Instead of strapping on alien devices and weighing ourselves down with cables and sensors, we should build remote sensing and perceptual intelligences into the environment. Instead of trying to recreate a sense of place by strapping video-phones and position/orientation sensors to our heads, we should strive to make as much of the real environment as possible responsive to our actions.

We have therefore chosen to build vision and audition systems to obtain the necessary detail of information about the user. We have specifically avoided solutions that require invasive methods, like special clothing, unnatural environments, or even radio microphones.

This paper describes a collection of technology and experiments aimed at investigating this domain of interactive spaces. Section 2 describes some our solutions to the non-invasive interface problem. Section 3 discusses some of the design challenges involved in applying these solutions to specific application domains with particular attention paid to the whole user: both their visual appearance, and the noises that they make.

2 Interface Technology

The ability to enter the virtual environment just by stepping into the sensing area is very important. The users do not have to spend time "suiting up," cleaning the apparatus, or untangling wires. Furthermore, social context is often important when using a virtual environment, whether it be for game playing or designing aircraft. In a head mounted display and glove environment, it is very difficult for a bystander to participate in the environment or offer advice on how to use the environment. With unencumbered interfaces, not only can the user see and hear a bystander, the bystander can easily take the user's place for a few seconds to illustrate functionality or refine the work that the original user was creating. This section describes the methods we use to create such systems.

2.1 The Interactive Space

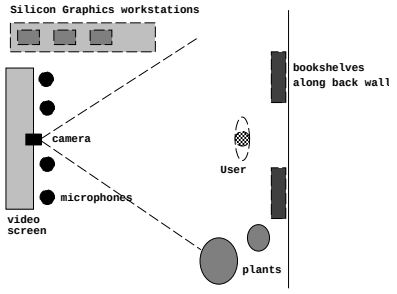

Figure 2: Interactive Virtual Environment hardware.

Figure 1: Netrek Collective Interface: the user issues audio and gestural information in conjunction.

Figure 2 demonstrates the basic components of an Interactive Space that occupies an entire room. We also refer to this kind of space as an Interactive Virtual Environment (IVE). The user interacts with the virtual environment in a room sized area (15'x17') whose only requirements are good, constant lighting and an unmoving background. A large projection screen (7'x10') allows the user to see the virtual environment, and a downward pointing wide-angle video camera mounted on top of the projection screen allows the system to track the user (see Section 2.2). A narrow-angle camera mounted on a pan-tilt head is also available for fine visual sensing. One or more Silicon Graphics computers are used to monitor the input devices in real-time.[10].

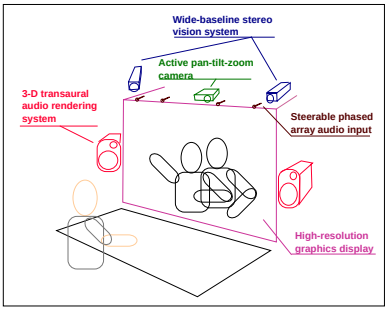

Figure 3: An Instrumented Desktop

Another kind of Interactive Space is the desktop. Our prototype desktop systems consist of a medium sized projection screen (4'x5') behind a small desk (2'x5'—See Figure 3). The space is instrumented with a wide-baseline stereo camera pair, an active camera, and a phased-array of microphones. This configuration allows the user to view virtual environments while sitting and working at a desk. Gesture and manipulation occur in the workspace defined by the screen and desktop. This sort of interactive space is better suited for detailed work.

2.2 Vision-based Blob Tracking

Applications such as unencumbered virtual reality interfaces, performance spaces, and information browsers all have in common the need to track and interpret human action. The first step in this process is identifying and tracking key features of the user's body in a robust, real-time, and non-intrusive way. We have chosen computer vision as one tool capable of solving this problem across many situations and application domains.

We have developed a real-time system called Pfinder[13] ("person finder") that substantially solves the problem for arbitrarily complex but single-person, fixed-camera situations(see Figure 4a). The system has been tested on thousands of people in several installations around the world, and has been found to perform quite reliably.[13]

Figure 4: Analysis of a user in the interactive space. Frame **(left)** is the video input (n.b. color image possibly shown here in greyscale for printing purposes), frame **(center)** shows the segmentation of the user into blobs, and frame **(right)** shows a 3-D model reconstructed from blob statistics alone (with contour shape ignored).

Pfinder is descended from a variety of interesting experiments in human-computer interface and computer mediated communication. Initial exploration into this space of applications was by Krueger [7], who showed that even 2-D binary vision processing of the human form can be used as an interesting interface. Pfinder goes well beyond these systems by providing a detailed level of analysis impossible with primitive binary vision.[13]

Pfinder uses a stochastic approach to detection and tracking of the human body using simple $2\frac{1}{2}$-D models. It incorporates *a priori* knowledge about people primarily to bootstrap itself and to recover from errors. This approach allows Pfinder to robustly track the body in real-time, as required by the constraints of human interface.[13]

Pfinder provides a modular interface to client applications. Several clients can be serviced in parallel, and clients can attach and detach without affecting the underlying vision routines. Pfinder performs some detection and classification of simple static hand and body poses. If Pfinder is given a camera model, it also back-projects the 2-D image information to produce 3-D position estimates using the assumption that a planar user is standing perpendicular to a planar floor (see Figure 4c)[13].

2.3 Stereo Vision

The monocular-Pfinder approach to vision generates the $2\frac{1}{2}$-D user model discussed above. That model is sufficient for many interactive tasks. However, some tasks do require more exact knowledge of body-part positions.

Our success at 2-D tracking motivated our investigation into recovering useful 3-D geometry from such qualitative, yet reliable, feature finders. We began by addressing the basic mathematical problem of estimating 3-D geometry from blob correspondences in displaced cameras. The relevant unknown 3-D geometry includes the shapes and motion of 3-D objects, and optionally the relative orientation of the cameras and the internal camera geometries. The observations consist of the corresponding 2-D blobs, which can in general be derived from any signal-based similarity metric.[1]

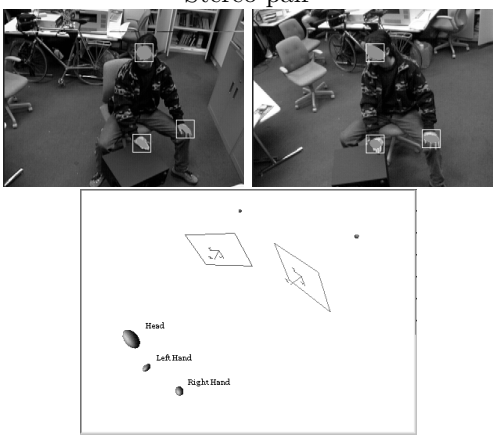

Figure 5: Real-time estimation of position, orientation, and shape of moving human head and hands.

We use this mathematical machinery to reconstruct 3-D hand/head shape and motion in real-time (30 frames per second) on a pair of SGI O2 workstations without any special-purpose hardware.

2.4 Physics-Based Models

The fact that people are embodied places powerful constraints on their motion. An appropriate model of this embodiment allows a perceptual system to separate the necessary aspects of motion from the purposeful aspects of motion. The necessary aspects are a result of physics, are predictable. The purposeful aspects are the direct result of a person attempting to express themselves through the motion of their bodies. By taking this one thoughtful step closer to the original intentions of the user, we open the door to better interfaces. Understanding embodiment is the key to perceiving expressive motion.

We have developed a real-time, fully-dynamic, 3-D person tracking system that is able to tolerate full (temporary) occlusions and whose performance is substan-

tially unaffected by the presence of multiple people. The system is driven by *blob features* as described above. These features are then probabilistically integrated into a fully-dynamic 3-D skeletal model, which in turn drives the 2-D feature tracking process by setting appropriate prior probabilities.

This framework has the ability to go beyond passive physics of the body by incorporating various patterns of control (which we call 'behaviors') that are *learned* from observing humans while they perform various tasks. Behaviors are defined as those aspects of the motion that cannot be explained by passive physics alone. In the untrained tracker these manifest as significant structure in the innovations process (the sequence of prediction errors). Learned models of this structure can be used to recognize and predict this purposeful aspect of human motion.[14]

2.5 Visually Guided Input Devices

Robust knowledge of body part position and body pose enables more than just gross gesture recognition. It provides boot-strapping information for other methods to determine more detailed information about the user. Electronically steer-able phased array microphones can use the head position information to reject environmental noise. This provides the signal-to-noise gain necessary for remote microphones to be useful for speech recognition techniques [2]. Active cameras can also take advantage of up-to-date information about body part position to make fine distinctions about facial expression, identity, or hand posture.[6]

2.6 Audio Perception

Recently, we have been moving away from commercial recognizers and working with the details of the audio signal. In the Self-Awear system, a wearable computer clusters dynamic models of audio and video features to find consistent patterns in the data. This allows the system to differentiate between environments that appear similar yet sound different (and vice versa) [5]. In the TOCO project, a robotic bird interacts with a user to learn the names of objects and their properties. The system has no prior knowledge of English except for phonemes, and the user speaks to it in complete sentences (e.g., "this is a red ball"). The system then uses consistent cooccurrences in the two modalities to determine what acoustic chunks are associated with what visual properties. As a result, it is able to successfully extract nouns and adjectives corresponding to object names, shapes, and colors [9].

In current work, we are trying to make use of prosodic information to understand the cues of natural speech, both in the audio (pitch, energy, timing) and the visual (head motions, expressions) domains. Whereas speech recognition has focused almost entirely on the dictation task, in which speech is spoken evenly and follows the rules of grammar, we are interested in the situations involving natural interactions between users or between a user and a machine. Though grammar no longer applies to this situation, the audio visual cues (changes in pitch, whether the head is facing the agent, etc.) should provide the necessary information to direct the receiver's understanding of speech[3].

3 Perceptive Spaces

Unencumbered interface technologies do not, by themselves, constitute an interface. It is important to see how they come together with the context of an application. This section describes several systems that have been built in our lab. as well as ongoing work. They illustrate a progression from early audio-visual interfaces employing a low amount of interaction between the modalities to current work on more complex modal integration.

3.1 SURVIVE

a. 

b. 

Figure 6: **(a)** The user environment for SURVIVE. **(b)** User playing with the virtual dog in the ALIVE space

The first smart space application to employ both audio and visual perception was the entertainment application SURVIVE (Simulated Urban Recreational Violence Interactive Virtual Environment). Figure 6a shows a user in SURVIVE.

The user holds a large (two-handed) toy gun, and moves around the IVE stage. Position on the stage is fed into Doom's directional velocity controls. The hand position features are used to drive Doom's rotational velocity control. We built a simple IIR filterbank to discriminate between the two sounds produced by the toy gun. The results of the matched-filter provides control over weapon changes and firing. The vision system is used to constrain the context for the audio processing by only operating when a user was detected on the IVE stage.

Although simplistic, this interface has some very important features: low lag, intuitive control strategy, and

a control abstraction well suited to the task. The interface processing requires little post-processing of the interface features, so it adds very little lag to the interface. Since many games have velocity-control interfaces, people adapt quickly to the control strategy because it meshes with their expectations about the game.

3.2 ALIVE

ALIVE combines autonomous agents with an interactive space. The user experiences the agents through a "magic-mirror" paradigm (including hamster-like creatures, a puppet, and a well-mannered dog—Figure 6b). The interactive space mirrors the real space on the other side of the projection display, and augments that reflected reality with the graphical representation of the agents and their world (including a water dish, partitions, and even a fire hydrant).

The "magic-mirror" model is attractive because it provides a set of domain constraints which are restrictive enough to allow simple vision routines to succeed, but is sufficiently unencumbered that is can be used by real people without training or a special apparatus.[8]

ALIVE employed a gesture-language that allowed the user to press buttons in the world or communicate wishes to the agents. ALIVE also employed audio perception. A commercial speech recognizer was used to turn speech events into commands for the agents. In this way speech provided a redundant modality to communicate with the agents.

Commercial speech recognizers require a very clean audio signal. It was critical to maintain the "hands-free" aspect of the interface, so we were unwilling to use a wireless headset or other such solution. Instead, we used a phased array of microphones that could be electronically steered to emphasize input from the user. The orientation of the steering was driven by the vision system, which could reliably track the user position [2]. In this way, the two modalities cooperated at the signal level.

3.3 City of News

City of News is an immersive, interactive web browser that makes use of people's strength remembering the surrounding 3D spatial layout. For instance, everyone can easily remember where most of the hundreds of objects in their house are located. We are also able to mentally reconstruct familiar places by use of landmarks, paths, and schematic overview mental maps. In comparison to our spatial memory, our ability to remember other sorts of information is greatly impoverished. City of News capitalizes on this ability by mapping the contents of URLs into a 3D graphical world projected on the large DESK screen. This virtual world is a dynamically growing urban landscape of information which anchors our perceptual flow of data to a cognitive map of a virtual place. Starting from a chosen "home page" - where home is finally associated with a physical space - our browser fetches and displays URLs so as to form skyscrapers and alleys of text and images through which the user can navigate. Following a link causes a new building to be raised in the district to which it belongs, conceptually, by the content it carries and content to be attached onto its "facade".

By mapping information to familiar places, which are virtually recreated, we stimulate association of content to geography. This spatial, urban-like, distribution of information facilitates navigation of large information databases, like the Internet, by providing the user with a cognitive spatial map of data distribution. This map is like an urban analogue to Yates' classical memory-palace information memorization technique.

To navigate this virtual 3D environment, users sit in front of the SMART DESK and use voice and hand gestures to explore or load new data. (Figure 7). Pointing to a link or saying "there" will load the new URL page. The user can scroll up and down a page by pointing up and down with either arm, or by saying "up/down". When a new building is raised and the corresponding content is loaded, the virtual camera of the 3D graphics world will automatically move to a new position in space that constitutes an ideal viewpoint for the current page. Side-pointing gestures, or saying "previous/next", allows to navigate along an information path back and forth. Both arms stretched to the side will show a full frontal view of a building and its contents. Both arms up drive the virtual camera up above the City and give an overall color-coded view of the urban information distribution. All the virtual camera movements are smooth interpolations between "camera anchors" that are invisibly dynamically loaded in the space as it grows. These anchors are like rail tracks which provide optimal viewpoints and constrained navigation so that the user is never lost in the virtual world.

The browser currently supports standard HTML with text, pictures and MPEG movies. City of News was successfully shown at the Ars Electronica 97 Festival as an invited presentation.

A phased array was not necessary here since the user was seated at a desk, but the coupling of the gesture and speech modalities was critical to making a robust, usable interface. The visual cues are used to activate the commercial speech system, thus avoiding false recognitions during speech not addressed to the system. Speech in turn is used to disambiguate gestures[12].

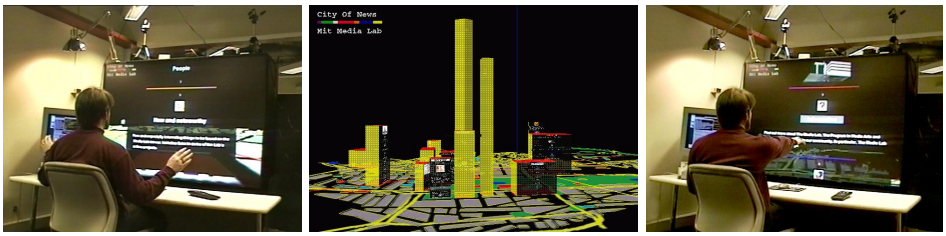

Figure 7: City of News.

3.4 The Perceptive Dance and Theater Stage

We have built an interactive stage for a single performer which allows us to coordinate and synchronize the performer's gestures, body movements, and speech, with projected images, graphics, expressive text, music, and sound. Our vision and auditory sensors endow digital media with perceptual intelligence, expressive and communicative abilities, similar to those of a human performer (Media Actors). Our work augments the expressive range of possibilities for performers and stretches the grammar of the traditional arts rather than suggesting ways and contexts to replace the embodied performer with a virtual one [11].

In Improvisational TheaterSpace, Figure 8a, we create a situation in which the human actor can be seen interacting with his own thoughts in the form of animated expressive text projected on stage. The text is just like another actor able to understand and synchronize its performance to its human partner's gestures, postures, tone of voice, and words. Expressive text, as well as images, extend the expressive grammar of theater by allowing the director to show more of the character's inner conflicts, contrasting action/thought moments, memories, worries, desires, in a way analogous to cinema. A pitch tracker is used to "understand" emphasis in the perfromer's acting, and its effects are amplified or contrasted by the expressive text projected on stage. The computer vision's feature tracker is then used to align the projection with the performer. Gesture recognition gives start/stop/motion cues to the Media Actors.

Improvisational Theater Space followed research on DanceSpace, a computer vision driven stage in which music and graphics are generated on the fly by the dancer's movements, Figure 8b.

3.5 Netrek Collective

Netrek is a game of conquest with a Star Trek motif. Netrek is very much a team-oriented game. Winning requires a team that works together as a unit. This fact, in particular, provides a rich set of interface opportunities ranging from low-level tactics to high-level strategy. The Netrek Collective is an example of our current work toward interfaces with more cross-modal integration.

The first version of the Netrek Collective, entitled *Ogg That There*, is intended to perform in a manner similar to the classic interface demo "Put That There"[4]. Imperative commands with a subject-verb-object grammar can be issued to individual units. These commands override the robots internal action-selection algorithm, causing the specified action to execute immediately. Objects can either be named explicitly, or referred to with deictic gestures combined with spoken demonstrative pronouns. Figure 1 depicts a user selecting a game object with a deictic gesture.

Much like City of News, deictics are the only form of gesture supported by *Ogg That There*. They are labeled by speech events, not actually recognized. The grammar is currently implemented as a finite state machine, and speech recognition is accomplished with an adaptive speech recognizer developed in the lab[9].

Ogg That There succeeded in solving many integration issues involved in coupling research systems to existing game code. Current work involves redesigning the interface to more accurately match the flexibility of the perceptual technologies, the pace of play, and the need for a game interface to be fluid and fun.

This will mean even richer interaction between gesture and speech. The biggest challenges in this work are: the integration of the significant game context with speech and gesture to provide a more robust and expressive interface. This will involve combining the perceptual tools discussed in Sections 2.4 and 2.6 with a dynamic constraint system linking these perceptual signals to the changing game context.

4 Conclusion

Throughout these projects, we have always tried to take advantage of the coupling of speech with other modalities. Our impression is that it is only by exploiting the connections between such domains that we can hope to construct truly natural interfaces. Though the various

a. 

b. 

Figure 8: **(a)** Improvisational performer Kristin Hall in the Perceptive Stage, at the MIT Media Lab, during the Digital Life Open House, on March 11, 1997. **(b)** Performer Jennifer DePalo, from Boston Conservatory, in DanceSpace, during rehersal.

pieces of our systems have become more complex over time, this philosophy continues to be an important factor in our continued work.

References

[1] Ali Azarbayejani and Alex Pentland. Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features. In *Proceedings of 13th ICPR*, Vienna, Austria, August 1996. IEEE Computer Society Press.

[2] Sumit Basu, Michael Casey, William Gardner, Ali Azarbayejani, and Alex Pentland. Vision-steered audio for interactive environments. In *Proceedings of IMAGE'COM 96*, Bordeaux, France, May 1996.

[3] Sumit Basu and Alex Pentland. Headset-free voicing detection and pitch tracking in noisy environments. Technical Report 503, MIT Media Lab Vision and Modeling Group, June 1999.

[4] R. A. Bolt. 'put-that-there': Voice and gesture at the graphics interface. In *Computer Graphics Proceedings, SIGGRAPH 1980,*, volume 14, pages 262–70, July 1980.

[5] Brian Clarkson and Alex Pentland. Unsupervised clustering of ambulatory audio and video. In *ICASSP'99*, 1999.

[6] T. Darrell, B. Moghaddam, and A. Pentland. Active face tracking and pose estimation in an interactive room. In *CVPR96*. IEEE Computer Society, 1996.

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

[8] Pattie Maes, Bruce Blumberg, Trevor Darrell, and Alex Pentland. The alive system: Full-body interaction with animated autonomous agents. *ACM Multimedia Systems*, 5:105–112, 1997.

[9] Deb Roy and Alex Pentland. "learning words from audio-visual input. In *Int. Conf. Spoken Language Processing*, volume 4, page 1279, Sydney, Australia, December 1998.

[10] Kenneth Russell, Thad Starner, and Alex Pentland. Unencumbered virtual environments. In *IJCAI-95 Workshop on Entertainment and AI/Alife*, 1995.

[11] Flavia Sparacino, Christopher Wren, Glorianna Davenport, and Alex Pentland. Augmented performance in dance and theater. In *International Dance and Technology 99*, ASU, Tempe, Arizona, February 1999.

[12] C. Wren, F. Sparacino, A. Azarbayejani, T. Darrell, James W. Davis, T. Starner, Kotani A, C. Chao, M. Hlavac, K. Russell, Aaron Bobick, and Pentland A. Perceptive spaces for performance and entertainment (revised). In *ATR Workshop on Virtual Communication Environments*, April 1998.

[13] Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.

[14] Christopher R. Wren and Alex P. Pentland. Dynamic models of human motion. In *Proceedings of FG'98*, Nara, Japan, April 1998. IEEE.