M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 398
November, 1996. Revised June 1998
Appears in: Presence: Teleoperators and Virtual Environments, Vol. 8, No. 4, August 1999, pp. 367-391.

The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment

Aaron F. Bobick, Stephen S. Intille, James W. Davis
Freedom Baird, Claudio S. Pinhanez
Lee W. Campbell, Yuri A. Ivanov, Arjan Schütte, Andrew Wilson

The MIT Media Laboratory
20 Ames Street
Cambridge, MA 02139
(kidsroom@media.mit.edu)

Abstract

The KidsRoom is a perceptually-based, interactive, narrative playspace for children. Images, music, narration, light, and sound effects are used to transform a normal child's bedroom into a fantasy land where children are guided through a reactive adventure story. The fully-automated system was designed with the following goals: (1) to keep the focus of user action and interaction in the physical, not virtual space; (2) to permit multiple, collaborating people to simultaneously engage in an interactive experience combining both real and virtual objects; (3) to use computer-vision algorithms to identify activity in the space without requiring the participants to wear any special clothing or devices; (4) to use narrative to constrain the perceptual recognition, and to use perceptual recognition to allow participants to drive the narrative; (5) to create a truly immersive and interactive room environment.

We believe the KidsRoom is the first multi-person, fully-automated, interactive, narrative environment ever constructed using non-encumbering sensors. This paper describes the KidsRoom, the technology that makes it work, and the issues that were raised during the system's development. [1]

I. Motivation and Background

1 Introduction

We are investigating the technologies required to build perceptually-based interactive and immersive spaces — spaces that respond to people's actions in a real, physical space by augmenting the environment with graphics, video, sound effects, light, music, and narration.

Using computer vision-based action recognition and other non-encumbering sensing technologies to interpret what people are doing in a space, a room can automatically provide entertaining feedback in a natural way by manipulating the physical environment. For example, a kitchen might use audio and video to guide its occupants through the preparation of a recipe; a cafe might observe how people are interacting and change lighting, video, and music to enliven the atmosphere; and, a child's bedroom might stimulate a child's imagination by using images and sound to transform itself into a fantasy world.

In this paper we detail our experience constructing and testing the KidsRoom, a fully-automated, interactive, narrative playspace for children. The space theatrically resembles a children's bedroom, complete with furniture including a movable bed. Under computer control and in response to the children's actions, the room uses two large back-projected video screens, four speakers, theatrical lighting, three video cameras, and a microphone to carry the children through a story. The KidsRoom experience was designed primarily for children ages six to ten years old, and it lasts ten to twelve minutes, depending upon how the participants act in the room. Throughout the story, children interact with objects in the room, with one another, and with virtual creatures projected onto the walls. The actions and interactions of the children drive the narrative action forward. Most importantly, the children are aware that the room is responsive.

The text is divided into three major sections, covering the motivation, implementation, and analysis of the project. We have included details at all levels of design ranging from broad considerations such as modeling the focus of attention of the user, to technical issues such as visual sensor integration, to the implication of implementation details. Our goal is that the lessons we learned from this project will be valuable to others constructing such perceptually-controlled, interactive spaces.

2 Project goals

The initial goal of our group was the construction of an environment that would demonstrate various computer vision technologies for the automatic recognition of action. As we developed the design criteria for this project it became clear that this effort was going to be an exploration and experiment in the design of interactive spaces. The goals that shaped our choice of domain were the following:

Action in physical space. Because our computer vision research focuses on the recognition of actions

---

[1] A demonstration of the project, including videos, images, and sounds from each part of the story is available at http://vismod.www.media.mit.edu/vismod/demos/kidsroom and complements the material presented here.

performed by humans, we require that the users be engaged in activity taking place in the *real physical environment*, not the virtual screen environment. Our goal was to augment a real space, stimulating the imagination using video, light, and sound, but not replacing the natural, real-world activity with which people are comfortable.

Vision-based remote sensing. Remote sensing permits unencumbered activity in the physical space, not requiring users to wear sensors, head-mounted displays, earphones, microphones, or specially-colored clothing. Also, computer vision tracking and action recognition techniques allow a person to easily enter and exit the room at any time naturally without troublesome sensor requirements.[2]

Multiple people. "Interactive" entertainment spaces are more engaging and social (and fun!) when one can play as part of a group. However, previous work in computer vision and fully-automated interactive spaces has primarily considered environments containing only one or at most two people. Our goal is to allow multiple people in the environment, interacting not only with the environment but also with each other. If unencumbered by head-mounted displays, people will naturally communicate with each other about the experience as it takes place, and they will watch and mimic one another's behavior.

Use of context. One of our research topics is the use of *context* to increase the reliability of vision-based sensing. Our goal is to build a system that is not only aware of the context of the situation (e.g. the current position in a story-line) but that also manipulates context by controlling much of the environment.

Presence, engagement, and imagination. We want an environment that is truly immersive, perceptually and cognitively engaging, and where participants do not need to ask for outside help. Furthermore, we want to create a *narrative* experience, where the story and action of the people with respect to the story are the primary focus. The experience should engage each user's imagination much like a children's story book; it is not necessarily required or desirable to provide a complete fanciful rendition of a virtual world. The experience should be compelling in the sense that the users should be more concerned with their own actions and behaviors than with how the interactive system works.

Children. We want the environment to be tailored to children. Davenport and Friedlander [10] and Druin and Perlin [14] both observed that adults visiting their interactive installations sometimes had difficulty immersing themselves in the narrative. Children already like to play with each other in real spaces enhanced by imaginary constructs (e.g. couches as caves) and can be easily motivated by supplementary imagery, sound and lighting.[3]

The final design of the KidsRoom was intended to achieve each of these goals. The idea of a children's playspace immediately addresses most of the concerns: a large encompassing room with multiple children being active in an engaging activity. The goal of exploiting and controlling context was accommodated by embedding the experience in a narrative environment, where there is a natural story-line to drive the situation.

3 Interaction in augmented environments

Bederson and Druin classify work on computer interactive interface systems into those that focus on building interfaces where information is superimposed on the physical world and those that embed information into the physical world itself [4]. The majority of research in the human computer interface and computer graphics fields has focused on systems of the first form, where a user must wear gear such as glove sensors, specially-colored clothing, or microphones. The KidsRoom is an example of systems of the second form, where the computer interface becomes unobtrusively embedded in the physical world itself. Such systems have been alternately termed "augmented environments," "immersive environments," "intelligent rooms" [31], "smart rooms" [23], and "interactive spaces."

One early example of augmenting a physical space was the "Media Room" project of Bolt and Negroponte [6]. Their system allowed a user sitting in a chair, ostensibly in his or her future living room, to interact with a screen by pointing and talking. Their goal, as is ours, was to augment spaces that we are comfortable with, but the technology available at that time required the use of body gear for sensing gesture and speech. Even today, most work in augmented environments requires cumbersome sensing and head-mounted display gear [1].

Research on physical, remotely sensed, interactive spaces began with Krueger's Videoplace system [20]. Krueger designed installations that explored many different modes of interaction, most of which entailed large body gesture. In one example, the user interacted with his or her own silhouette on video screens.

The ALIVE project improved on Krueger's system by replacing special blue-screening hardware with computer vision algorithms that can track the position and gestures of a single person moving in front of an arbitrarily complex, static background [22]. A single user can interact with virtual creatures by watching his or her own image superimposed with behavior-based creatures on a large video wall. The user must orient towards the video wall to observe the interesting action.

The "Intelligent Room" system consists of several cameras and two large screens in a small room [31]. A single

---

[2] Some of the "direct sensing" tasks could be implemented using other devices such as micro-switches to detect presence of a person on a bed. Because our focus is on computer vision we did not employ such devices. However, even if a room is densely wired with sensing devices, the difficult problem of *understanding what is happening* in the space still needs to be addressed, and that is a fundamental component of our research on understanding action in the vision domain [5].

[3] We also note that children are more forgiving of the small glitches in timing, animations, and recognition certain to be present in such an experimental facility; our hope is that children will focus more on having fun in the space than on figuring out how it works or how to break it.

person is tracked using computer vision, and simple pointing gestures are recovered. Users can utter a lexicon of about 25 sentences into a lapel microphone. The computer will understand instructions such as "Computer, what is the weather here?" and will use the city that the person is pointing to on one of two large screens to retrieve weather forecasting information that is then displayed on the other screen. The room is controlled by a distributed agent-based architecture [8]. The goal of the project is to remove the computer from the human-computer interface.

An alternative approach to the design of interactive spaces is to mediate computer interaction through the manipulation of real, physical objects. Druin, for instance, constructed a stuffed-animal doll called Noobie [13]. Children interacted with Noobie by squeezing the doll's limbs and watching a display embedded in its belly. Instead of bringing children to the virtual space on the screen and forcing interaction with special devices, the interface was brought into the world of the children and embedded into devices with which they were already comfortable (also see [33, 15]).

Druin and Perlin set out to construct immersive physical environments for adults that responded to movement and touch in real physical spaces [14]. They supplemented a real environment with a simple narrative and in 1993 debuted an interactive installation with three stories: one humorous story about baby sitting, another more serious narrative about heaven and hell, and a final murder-mystery scenario. The system used computer control of lighting, sound, video, and physical props and sensor embedded in objects. Interestingly, Druin and Perlin comment that some adults were confused with the whole idea of an immersive experience. One participant commented, "I didn't think I should touch anything. You know, Mom always said, Do not touch!"

Narrative in interactive physical spaces was further explored by Davenport and Friedlander [10]. They wanted people to "feel as though they were walking through a computer monitor into a magic landscape." They constructed a four-world, *human-controlled* installation where the narrative was "actualized by the transformative actions of the visitor moving through it." The room used light, sound, video, and computer displays. Each person in the space had a human guide, another "user" of the system, outside the space communicating with him or her via computers.

A variety of artistic experiments have been undertaken involving computerized spaces. A review of this work is beyond the scope of this paper, but references and a critical analysis of some of the experiments are available [21, 26]. A notable installation is Masaki Fujihata's *Beyond Pages*, featuring a virtual book whose illustrations of objects such as a lamp and door react to the user's gestures. Artists Christa Sommerer, Laurent Mignonneau, and Naoko Tosa have integrated computer vision, computer graphics, and emotion and speech recognition techniques [30, 32].

Bederson and Druin believe, as do we, that the best immersive physical environments will have multi-modal inputs and outputs [4]. Increasing computational speed is now making it possible to explore domains like immersive office environments [35, 18], living spaces, and theater performances [25]. The two major obstacles to building fully-automated reactive spaces are (1) finding practical, computationally-feasible sensing modalities that can be used to understand a variety of different types of human action and interaction and (2) developing a computationally-feasible control mechanism and inter-communication architecture for coordinating perceptual input, narrative control, and perceptual output systems. Both these goals require that we further our understanding of how we represent actions, interaction, time, and story.

4 Recognition of action

Most virtual reality systems map a given configuration of the sensor outputs directly to some system response. For instance, in the ALIVE system, the position of a person's hands and head are estimated using computer vision, and the relative position of these objects is used to determine if a person is making a gesture [36].

The KidsRoom moves beyond just measurement of position towards recognition of *action* using measurement and context. Although many of the mechanisms used are simple, the KidsRoom combines the sensor outputs with contextual information provided by the story to recognize over a dozen simple individual and group actions in specific contexts. Examples include moving through a forest in a group, rowing a boat, and dancing with a monster. These recognized actions drive the story and control the narrative.

Throughout this paper, we will lump all types of action recognition together. However, within the computer vision community, researchers are developing a taxonomy of action based on the computational representations and methods required to understand each action type (e.g. a taxonomy of movement, activity, and action [5]). Many actions of interest require that contextual knowledge be used for recognition in addition to sensed motion and position information. In simple contexts, direct measurement of body position can sometimes be used to recognize activity. However, as the complexity of an environment increases, many different measurements may correspond to the same actions, or, depending on the context, the same measurement may correspond to different actions. Stronger contextual constraints are required to extract action labels from perceptual measurements. The environment of the KidsRoom is rich enough to begin to explore some context-sensitive recognition tasks.

II. Implementation and Experience

5 The Playspace

The KidsRoom theatrically re-creates a child's bedroom. The space is 24 by 18 feet with a wire-grid ceiling 27 feet high. Two of the bedroom walls resemble real walls of a child's room, complete with real furniture and decoration. The other two walls are large video projection screens, where images are back-projected from outside of the room. Behind the screens is a computer cluster with six machines that automatically control the room. Computer-controlled theatrical colored lights on the ceiling illuminate the space. Four speakers, one on each wall, project sound effects and music into the space. Finally, there are four video cameras and one microphone installed. Figure 1 shows a view of the complete KidsRoom installation.

The room contains several pieces of real furniture including a movable bed, which is used throughout the story. Because the other furniture is not explicitly tracked by the

Figure 1: The KidsRoom is a 24 by 18 foot space constructed in our lab. Two walls resemble the walls in a real children's room, complete with posters and windows. The other two walls are large back-projection screens. Computer-controlled lighting sits on a grid suspended above the space. The door to the space, where all room participants enter and exit, is pictured in the leftmost corner of the room.

computer vision system, it is sealed shut and fastened to the ground. Four colored rugs with animal drawings and simulated stone markers on the floor are used as reference points during the narrative. Colored cinder blocks on the floor prevent enthusiastic children from pushing the bed through the screens. Figure 2 shows the layout of the room's interior.

Figure 3 shows the four camera views. Camera one is the top view, which is used for tracking people and for some motion detection. Cameras two and three are used to recognize body movements when children are standing on the green and red rugs, respectively. Finally, camera four provides a spectator view of most of the room and the two projection screens; this view is displayed to spectators outside the space and provides video documentation. In addition to the visual input, a single microphone is in the space that is used to detect the loudness of shouts.

The room has five types of output for motivating participants: video, music, recorded voice narration, sound effects, and lighting. Still-frame video animation is projected on the two walls. Voices of the narrator and monsters, as well as other sound effects, are directionally controlled using the four speakers and appear to come from particular regions of the room. Some sound effects, such as monster growling and boat crashes, are particularly loud and can vibrate the floor, providing visceral input. As we discuss later, lighting must remain constant when the vision algorithms are operating; however, because the story can be used to determine when vision is and is not required, it is possible to use lighting changes and colored lighting to mark important transitions.

Six computers power the KidsRoom. One SGI Indy R5000 workstation is used for tracking people and the bed, playing sound effects, and MIDI control of light output. A second SGI Indy R5000 workstation is used for action recognition from cameras one and two, sending MIDI music commands to the music computer, and amplitude audio detection. A third SGI Indy workstation is used for action recognition from camera three. Two DEC AlphaStations are used for displaying still-frame animations, one per screen. One of the AlphaStations also runs the room's control process. A Macintosh is used for running Studio Pro MIDI software connected to a Korg 5R/W synthesizer.[4] Finally, assorted video, lighting, and sound equipment are required to complete the installation.[5]

6 The Story

The KidsRoom guides children through an interactive, imaginative adventure. Inspired by famous children's stories in which children are transported from their bedrooms to magical places (e.g. [2, 34, 27]), the story begins in a child's bedroom and progresses through three other worlds. We will describe the last world in detail, to give the reader a feel for the story, it's characters, and the interactive responsiveness of the entire system.

---

[4] We used the hardware we had available, but current PCs equipped with real-time video digitizers would suffice.

[5] Includes two high-resolution video projectors and wall-sized screens, 4 Sony HandyCam color video cameras, 4 speakers and a 12-channel 4-output mixer and amplifier, one microphone, 14 lights (11 white, 3 colored), a MIDI-based light board controller, and video distribution amplifiers.

Figure 2: The KidsRoom is furnished like a real children's room, complete with furniture, decorations, and a movable bed. Rugs and stone-like markers are used throughout the narrative. Four speakers project sound into the space. Colored cinder blocks at the base of the large projection screens protect the screens (which make up the right and bottom walls in this image) from the movable bed. The square on the floor in the bottom left marks the "door," where people enter and exit the space.

Camera 1

Camera 3