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

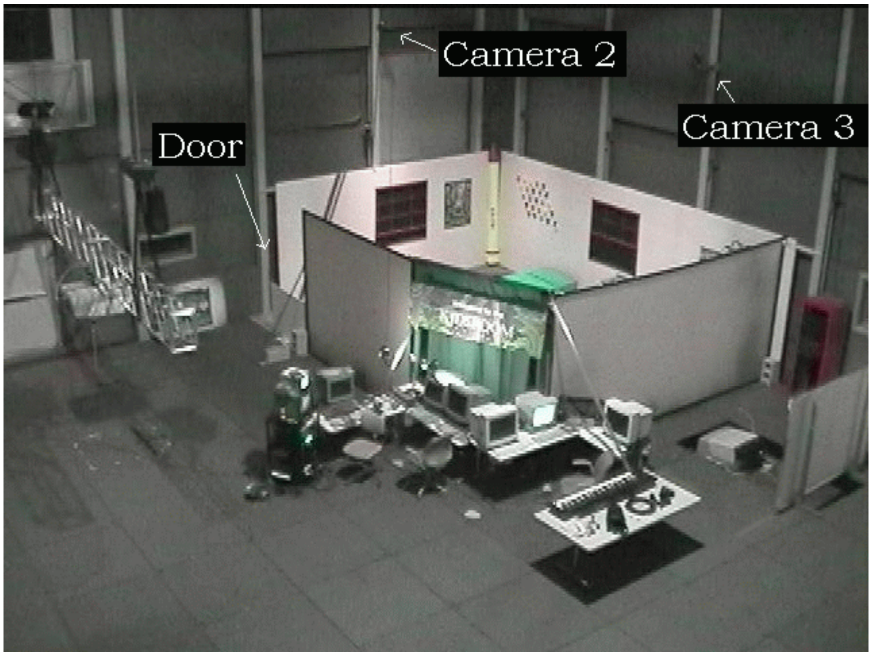


Figure 1: The KidsRoom is a 24 by 18 foot space constructed in our lab. Two walls resemble the walls in a real children's room, complete with posters and windows. The other two walls are large back-projection screens. Computer-controlled lighting sits on a grid suspended above the space. The door to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

Camera 2

Camera 4

Figure 3: Three cameras overlooking the KidsRoom are used for computer vision analysis of the scene. Camera 1 is used for tracking the people and the bed in the space and also for detecting motion during particular parts of the story. Cameras 2 and 3 are used to recognize actions performed by people standing on the red and green rugs. Camera 4 is used to provide a view of most of the room and the screens for spectators outside the space.

We note there that the primary story is a traditional linear narrative, as opposed to the (typically weakly) non-linear branching story-lines found in many multi-media presentations. Individual responses made by the room are reactive and in that sense non-linear. As we will discuss in the analysis section, we needed a strong narrative to motivate group behavior and to provide sufficient context. We will argue that this linear structure in no way reduced the interactivity because the pacing and individual reactions of the room are completely determined by the participants.

The only instruction given to the children prior to entering the room was that this was a magic room, but that to transform the room they needed to learn the magic password. To learn the password they should try "asking the furniture."

6.1 The bedroom world

Children enter the KidsRoom one at a time through the "door" in one corner of the room. The tracking algorithm

Figure 4: When participants first enter the space, the projection screens display the walls of a bedroom.

attends to this region, checking for people entering and exiting the space. Whimsical, curious music plays softly, and the projection walls display scenes from a bedroom, as shown in Figure 4. When at least one child approaches some piece of furniture (e.g. the blue desk or the green frog rug), each of which has a distinct personality, the furniture speaks and a scavenger hunt for the magic word commences, as shown in Figure 5a. For example, a clothes trunk says (with suitable accent), "Aye, matey, I'm the pirate chest. I don't know the magic word, but the frog on the rug might know." The children then run to the rug with the frog painted on it, and the rug frog seemingly speaks, sending them to yet another piece of furniture. The system randomly selects the ordering so that the interaction is slightly different each time the system is run.

Even a situation as simple as the bedroom requires handling contingencies. If the children get confused and do not go to the correct place, the system will eventually respond by having some furniture character call the children over, "Hey, over here, it's me, the yellow shelf!"

This game continues for a few iterations, usually with running, screaming, laughing children. Eventually, one piece of furniture knows the randomly-chosen magic word (e.g. "skullduggery") and reveals it to the children. As soon as the password is disclosed, all of the furniture start chanting the word loudly, ensuring the children hear and remember it. A mother's voice soon breaks in, silencing the furniture voices, and telling the kids to stop making noise and to go to bed. When they do, the lights drop down, and a spot on one wall is highlighted, which contains the image of a stuffed monster doll. The monster starts blinking and speaks, asking the children to loudly yell the magic word to go on a big adventure: *"On the count of three yell the magic word: One, two, three, . . ."*

6.2 The forest world

After the kids yell the magic word loudly,[6] the room darkens and the transformation occurs. Images on the projection walls gradually fade to images of a cartoon fantasy forest land, and colored, flashing lights combined with mysterious music play during the transformation, as shown in Figure 6. Simultaneously, a grandfatherly-voiced narrator — the voice and personality of the room — says, *"Welcome to the KidsRoom. It's not what it seems. What you might see here are things dreamt in your dreams."* The narrator always speaks in couplets, enhancing the experience of being immersed in a children's story book.

As the lights come up, the narrator tells the children that they are in the "Forest Deep," that monsters are near, and that they must follow the path to the river. A stone-path is marked on the floor of the room and the children quickly realize that is the path they should follow. The room provides encouraging narration if they do not do so and instructs them to to stay in a group and to remain on that path. If they deviate from the path "hints" are given to induce the behavior. Hints are loudly whispered suggestions made in a soft female voice to provide additional instruction when needed; their power will be discussed later in the analysis section.

As they traverse the path (moving around the room several times), monsters are heard growling from afar. The narrator warns *"The magic bed is now a tree. Hide behind where monsters cannot see."* When the children move behind the bed, the monsters stop growling, and the children continue on the path (see Figure 5b). If they do not hide, the loud growling intensifies and a different narration encourages the kids to get behind the bed. After a short walk, the children reach the river world.

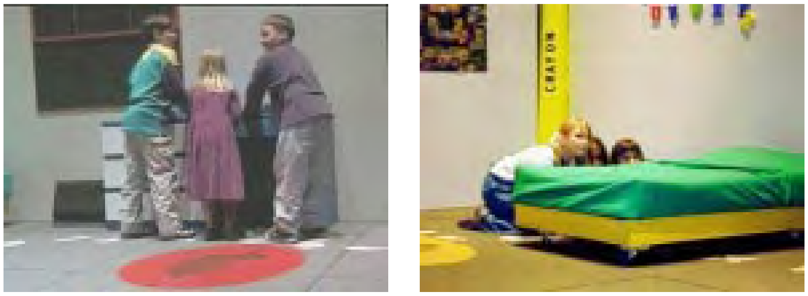
6.3 The river world

As the narrator announces *"You've certainly managed a glorious act. You've arrived at the river and you're still intact,"* images of a river dissolve onto the two screens. One view shows the river progressing forward (see Figure 7) and the other view shows the sideways moving riverbank. In the river world, the children are told the "magic bed" is now a boat. They are encouraged to push the bed to the center of the room and "jump inside" by climbing on top. When the children start making rowing motions, the river images start moving. If someone gets off the boat, a splashing sound is heard. The narrator then shouts a "passenger overboard" couplet and encourages the child to get back on the bed.

Soon, log and rock obstacles appear in the path of the boat, as shown in Figure 7a. As instructed by the narrator, the children must engage in collaborative rowing (making rowing motions on the correct side of the bed) to avoid the obstacles. If they successfully navigate the obstacle, a female voice softly whispers hints such as "Nice job." When they do not, they hear a loud crashing sound, often motivating the kids to physically play-act a crash as in Figure 7b, and they receive some whispered hints about

---

[6] There is no speech recognition capability, with only the volume of sound being measured. In no run of the KidsRoom did the children ever yell anything but the magic word, illustrating how a compelling narrative will constrain behavior.



(a) (b)

Figure 5: (a) Children are told only to "ask the furniture for the magic word." (b) During the forest world, children must hide behind the bed to stop the loud growling of distant monsters.

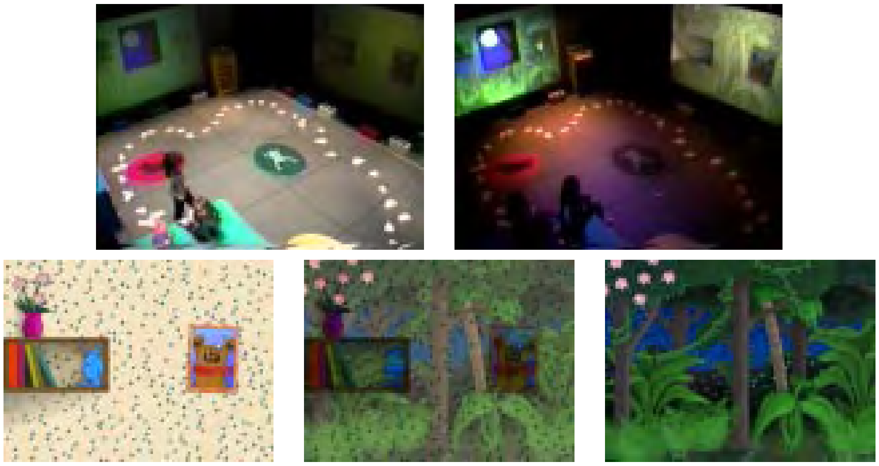


Figure 6: Lighting effects are used to mark special transitions in the story. Here colored lights flash, transitional music plays, and the screens gradually fade from bedroom to forest as the narrator welcomes the children to the forest world. Graphics are simple story-book animations. However, combined with music, narration, and lighting effects, the transformation captures the attention of people in the room and gives the room a somewhat magical quality.

how to avoid the obstacles.[7]

Eventually, the image of a shore appears, and the children are instructed to "land the boat" by pushing the bed towards the "shoreline" on the screen. As they do, the system produces loud grinding noises as if the bed is being pushed onto the sand. Suddenly, the mother's voice is heard in the distance again telling the children to stop moving furniture and to put the bed back where they found it. If necessary, additional hints are provided until the bed is returned to approximately its original position. At this point the children have reached the monster world.

6.4 The monster world

Forest images displayed on the two screens, tense forest music playing in the background, and jungle sounds like twigs cracking and owls hooting announce the arrival to the monster land. To give a feel for the interaction we present this world as an annotated dialog. The narrator speaks in a comforting, deep, somewhat mischievous voice.

> Narrator: "*Welcome to Monster Land. It's a great spot. Time to have fun, ready or not.*"

Monsters are heard growling softly, but cannot yet be seen. Children are often huddled on the bed waiting expectantly. Suddenly, there are loud roars and the monsters appear on the two screens. The monsters, shown in Figure 8a-b, are larger than the children and have a friendly, goofy, cartoon look. As they continue to growl, the room speaks and suggests that if the kids yell, the monsters might be quiet. The kids, in unison, yell. If the shout is loud, the story continues. If not, the room responds with encouragement.

> Narrator: "*Get those monsters back in line. Try that shout one more time!*"

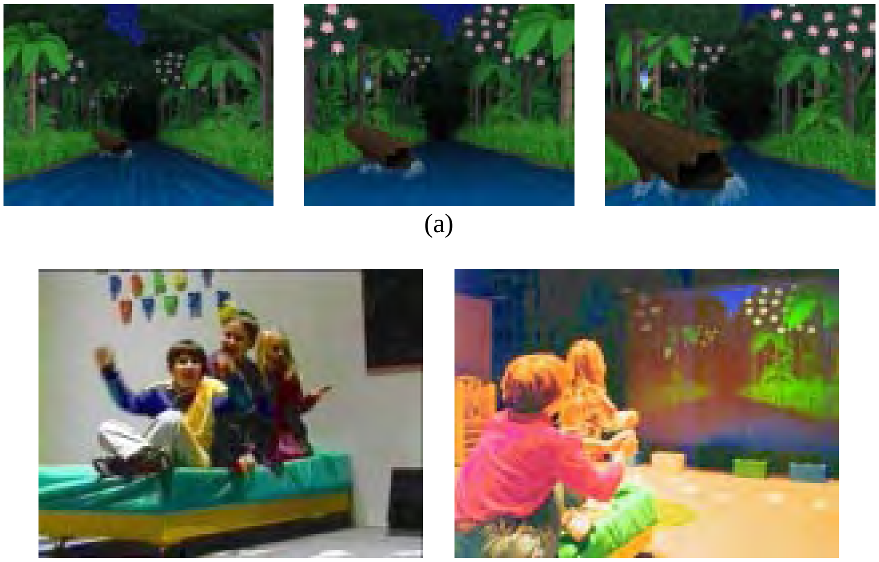
If the kids still do not scream, the room responds:

> Narrator: "*Well, I can't say* that *was a very loud shout, but perhaps the monsters will figure it out.*"

Either way, the monsters stop growling and show surprise for a moment. Energetic music starts to play.

---

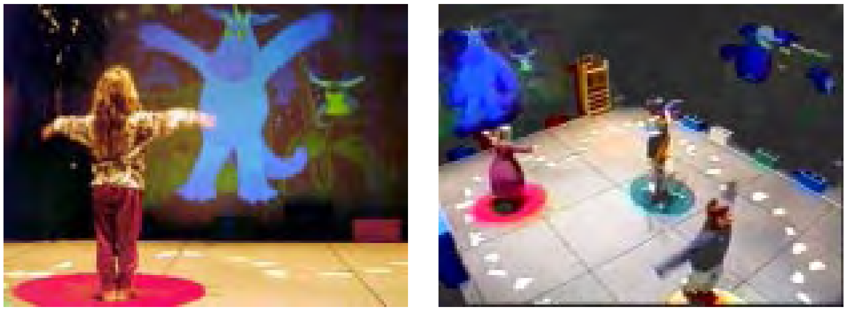
[7] We learned by experience that children prefer to crash and typically paddle *towards* the obstacles!



(a)

(b) (c)

Figure 7: (a) The speed of the boat is controlled by how vigorously people on the bed are rowing. If everyone stops making motions, the boat imagery will stop moving forward. Obstacles such as the log in this series of images approach the boat. They can be avoided by rowing strongly on the appropriate side of the bed (i.e. if the log is on the the left as shown, then row on the left-hand side of the bed). (b) Audio feedback such as loud crashes and narration signal when obstacles have been hit or avoided. Crashes tend to evoke expressive responses from children and subsequently more enthusiastic rowing. (c) A child and mother row the boat together.



(a) (b)

Figure 8: (a) A child dancing with a monster. (b) Children spinning during the monster dance.

> Narrator: *"The monsters invite you to shimmy and dance. Go stand on a rug and you'll get your chance."*

When the children are on rugs (sometimes prompted multiple times in different ways), the room continues. The vision algorithm used in this world requires there be only one child on each rug to get a non-occluded view of each participant. Therefore, if the system detects multiple people on a rug, one of the monsters in the story responds in his raspy voice:

> Monster 2: *"Hey, only one kid per rug please, so's we can see what's goin' on."*

Throughout this section of the story, the system detects when children get off of a rug and the characters in the story respond accordingly.

> Monster 2: *"Hey, be sure to stay on your rug there, thanks a lot."*

When each rug has a single child on it, the monsters begin to teach the children four dance moves:

> Monster 1: *"Hey, I'm going to do a crouch. You watch me first, then you try it. To do it right, crouch down and touch your toes."*

The monster, represented using still-frame animation as illustrated in Figure 9 does the move, and then says, *"Your* *turn!"* The system watches the children on two of the rugs.[8] When a child is observed having done the crouch move, the monsters respond with positive comments:

> Monster 1: *"Yo! Kid on the red rug, you dance like a pro!"*

The monsters continue, teaching the children three other moves: throwing the arms to make a 'Y', a flapping move with arms extended, and a spinning move also with arms out.

Once the children know the moves, the music changes and becomes more energetic.

> Narrator: *"Now that you know the monster moves, see if they can catch your grooves! The monsters will be able to copy if you do the moves, but don't be sloppy."*

The system will now respond to any of the four moves made by the children on the front two rugs. The monsters will mimic a move that a child performs. If the system has a high confidence that it knows which move the child is performing then the monster also offers spoken confirmation such as "Awesome flaps." If the children stop doing recognizable moves, the monster characters choose moves themselves and a whispered hint like, "Try a flap, spin, 'Y', or crouch" is heard.

This dance phase continues for a few minutes (see Figure 8b). Then the music grows louder and faster:

> Narrator: *"Now it's time to do your own dance. Let's see you dance and boogie all around the room!"*

The monsters, yelling kudos such as "Feel the groove!" and "Shake that funky thing!", do a variety of monster moves, including some new ones. The music's character, tempo and volume cause the children to run around the room and do new dance moves. Suddenly, the mother's voice is heard again, and all music and sound abruptly end:

> Mom: *"I told you kids to go to bed, and I mean business."*

If the children all get on the bed, the story moves on. However, if not, the mother character continues to encourage the desired action.

> Mom: *"I'm not fooling around. Get on that bed. All of you!"*

As soon as everyone's back on the bed, the monsters respond to the end of the scene:

> Monster 1:*"Hey y'all, let's get quiet, and next time you come back, we'll have a rockin' good time."*

The lights drop down and colored transition lighting is used. Transformational music plays, as the monsters say goodbye, "Take it easy! Bye bye!" The forest fades back to the room, and as the lights slowly come up:

> Narrator: *"Thanks for our Monster Land visit with you. We've enjoyed this wild dream, and we trust you did too."*

Exit music plays as the children leave the space.

7 Perceptual Technology

As discussed in section 4, in the KidsRoom we both *measure* the position and movement of multiple, interacting people and then use that data and contextual information to *recognize* action using vision-based perceptual algorithms. This section briefly describes the perceptual methods used by the KidsRoom. Discussion of the difficulties we encountered with each method are deferred until the analysis section.

7.1 Object Tracking

Most immersive environments will need to keep track of the positions of people and objects in the space. In the KidsRoom, we track the positions of up to four people and the movable bed. Some worlds, like the bedroom world, use positional information to know whether people are near certain objects: the pieces of furniture speak only when a person is near. The positional information is also used to keep track of whether people are in a group and whether they are moving or not. Most importantly, position information is used to create known contexts for other vision processes by ensuring that people are in expected regions of the room.

The KidsRoom tracking algorithm uses the overhead camera view of the space, which minimizes the possibility of one object occluding another. Further, lighting is assumed to remain constant during the time that the tracker is running. Our room lighting is designed to minimize brightness variation across the scene, but in practice an object's observed color and brightness can significantly change as it moves about.

Background subtraction is used to segment objects from the background, and the foreground pixels are clustered into 2D blob regions. The algorithm then maps each person known to be in the room with a blob in the incoming image frame. When blobs merge due to proximity of two or more children, the system maps more than one person to a given blob. The system uses color, velocity estimation, and size information to disambiguate the match when the blobs later separate. It is important for the algorithm to keep track of how many people are in the room, which is achieved by having everyone in the room enter and exit through a "door" region.[9] The context-sensitive, non-rigid object tracking method is fully described and tested elsewhere [17].

Figure 10 shows the image view used for tracking and the output of the tracking system. Each person is represented

---

[8] The children on the blue and yellow rugs near the back of the room are tracked but their actions are not analyzed by the system, because performing recognition on those rugs would require additional cameras and computers not at our disposal. Children are only on these rugs when more than two people are in the room, and in those cases the children are clearly aware of the interaction between the monsters and their playmates.

[9] During scene transitions, the lighting varies and the vision system is disabled. The story and timing of the narrative are designed such that nobody would exit during a transition and nobody ever did; a human gate-keeper prevented people from entering during those times. When lighting stabilized, therefore, the system knew the number of people in the room on the bed and could initialize the tracker accordingly.

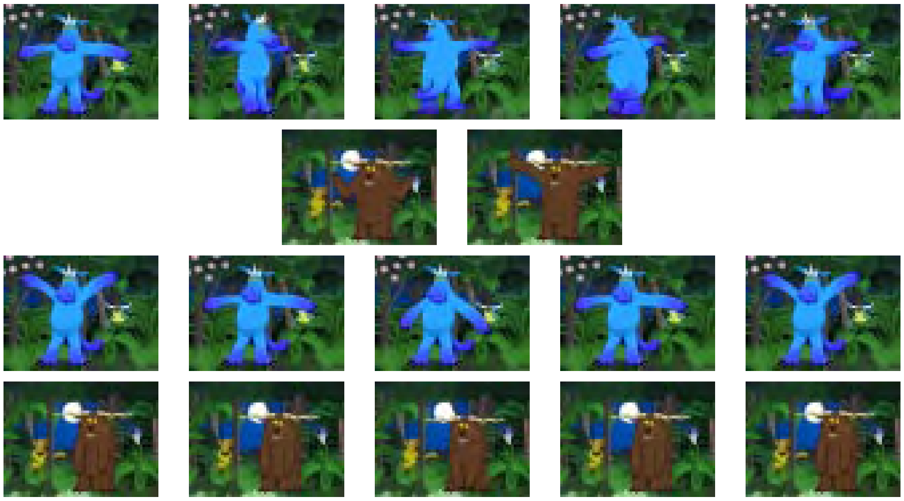


Figure 9: The dance moves are taught to the children by the monsters using still-frame animation. The first sequence shows one monster doing the spin move: "Put your arms out and spin around like a top." The second sequence shows another monster doing the 'Y' pose: "Throw your arms up and make a 'Y'." The third sequence shows "Flap your arms like a bird," and the last sequence shows "Crouch down and touch your toes."

by the rectangle bounding his or her blob. The box in the lower left corner represents the "door" region of the room, where people can enter and exit. The tracking algorithm is not limited to four people, but the KidsRoom narrative was designed for a maximum of four participants.

7.2 Movement detection

Earlier we made the distinction between measuring movement and recognizing action. A strongly constraining context, however, can allow inference of action directly from movement. For example, in the river world measurements of motion energy used in conjunction with contextual knowledge are employed to recognize participant's rowing actions. The amount of motion on each side of the bed is used by the control program to decide if the boat is moving (i.e. passengers are "rowing") and if the people have avoided obstacles in the river by rowing vigorously on the correct side of the boat.

The rowing detection algorithm presumes that everyone is "inside the boat" — all on the bed. The narrative encourages participants to establish and maintain this context (e.g. "Tuck your hands and feet right in, the hungry sharks are eager to sin."), and a simple vision algorithm based upon the size of the bed is used to confirm that the context is in effect achieved. When the blob size is about right, everyone is assumed "in the boat" and the bed orientation is computed.

Once the system knows everyone is on the bed and knows how the bed is oriented, it can use a simple test to check if there is more rowing on the left or right side. The algorithm computes the pixel-by-pixel difference between consecutive video frames. If someone moves quickly, a large difference between frames is detected. The difference over a region is the rowing energy, which is measured on each side of the bed and scaled by the number of people in the boat. The control program then uses these energy measures to detect whether or not people are rowing and on which side most of the rowing is occurring. Figure 11 shows the output of the system when a person is rowing on the left, right, and both sides of the bed.

7.3 Action recognition in the monster world

More sophisticated motion analysis is used during the dance segment of the monster world to recognize the four actions of "making a 'Y'," crouching, flapping and spinning. We chose these moves for several reasons: they are fun, natural gestures for children; they are easy to describe and animate using still-frame animation; they are easy to repeat in about the same way each time; and they allow us to demonstrate a few different recognition techniques using computer vision.

Each of the real-time approaches for recognition described below are run in parallel as the kids perform dance moves. The vision system reports which moves it thinks is being performed, as well as its confidence in that assignment. All of the vision processes use background subtracted images which contain only a silhouette of the person. They also require a training phase prior to run-time when the action models are constructed.

7.3.1 General dynamics

The first and simplest technique, for detecting crouching, uses the size of the background-difference blob. Once in the monster world, the "standing" blob shape for a person is initialized as soon as the person moves onto the rug. The blob shape, which is modeled using an ellipse matched to the blob data, is compared at each time with the "standing" model. If the elongation of the blob reduces significantly, the algorithm will signal that a crouch has taken place. Figure 12b shows a person's image blob and the ellipse model for standing and crouching positions.[10]

---

[10] For all moves, the control system ignores the move reported by the vision system if the tracking has indicated the person is not

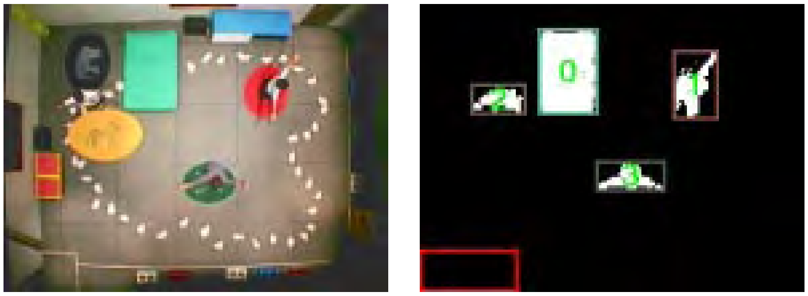


Figure 10: The left image shows a view with three people in the room from the overhead camera used for tracking. The right image shows the output of the tracking system, which is described and evaluated elsewhere [17]. All three people and the bed are being tracked as they move about. The box in the lower left denotes the room's door region, where all objects must enter and exit. The KidsRoom tracks up to four people and the bed.

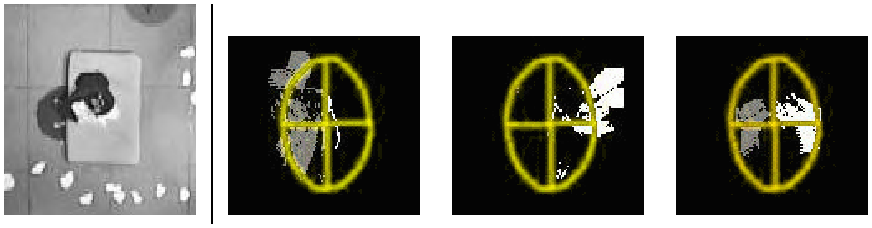


Figure 11: These images show the motion energy that is detected from the overhead camera when a person is "rowing" as they sit on top of the bed. The ellipse represents the position and orientation of the bed, extracted by the system; the colored pixels indicate where the system detected motion. The left, middle, and right images show rowing on the left, right, and both sides of the boat, respectively. The amount of movement at any time is compared with the maximum movement detected so far to compute how vigorously people are rowing and on which side of the bed.

7.3.2 Pose recognition

The second recognition technique uses the shape of the person's background-subtracted blob to identify when the person's arms are raised up in the air in a 'Y'. Here we use a pattern recognition approach to classify the background subtracted images of the person. Moment-based shape features [16] are computed from the the blob images like those shown in Figure 12d and are statistically compared to a database of training examples of people "making a 'Y'."

7.3.3 Movement recognition

The last technique used to recognize monster moves uses recognition of *motion templates* [11]. In this method, successive video frames of the background subtracted images of the people are temporally integrated to yield a "temporal template" of the action. Templates for the flap and spin moves are shown in Figure 13. These template descriptions represent the movement over some time interval with a single vector-valued image. The range of duration of integration is determined by training examples of the actions. A statistical moment-based description of the action template is then used to match to a database of examples of the moves.

7.4 Event Detection

In addition to recognizing large body motions of individuals, most immersive environments need to be able to detect other "events" if they are to provide interesting, reactive feedback. For example, the KidsRoom uses the output of the the tracker to answer questions such as "Is everyone in a group?", "Is everyone on the bed?", "Is everyone on the path?", "Is everyone moving around the path or standing still?", and "Is someone near a particular object?" The KidsRoom uses straightforward methods to compute answers to these questions. The "in-a-group" detector receives the position of each person from the vision tracker and validates that every person is within some predetermined distance of another person.

8 Story Control Technologies

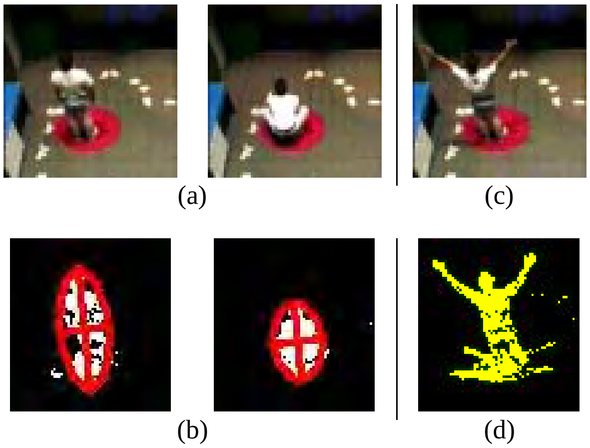
In addition to the perceptual input technology, the KidsRoom has a narrative control program, a lighting control program, midi music control programs, and networking protocols.

8.1 Narrative control

The narrative control program of the KidsRoom queries the sensor programs for information about what is happening in the room at a given time and then determines how the room responds so that participants are guided through the narrative. For example, when someone enters the room the system must start tracking the person and the control program must immediately learn of the person's presence. Similarly, if everyone leaves the room, the story must freeze at its current point instead of continuing on as if there were still participants. The main control program is an event

---

on the rug.


(a) (c)

(b) (d)

Figure 12: (a) A person performing a crouch move. (b) A person's background difference blob. Overlayed on top is an ellipse model of the blob. The first image shows a person in the standing position. The second shows the same person crouching. The difference in elongation of the ellipse model is used to detect crouching movement. (c) A person performing a 'Y' move. (d) The blob image used to detect the 'Y' move. This binary image is matched to a set of models of Y moves using moment-based shape-features.

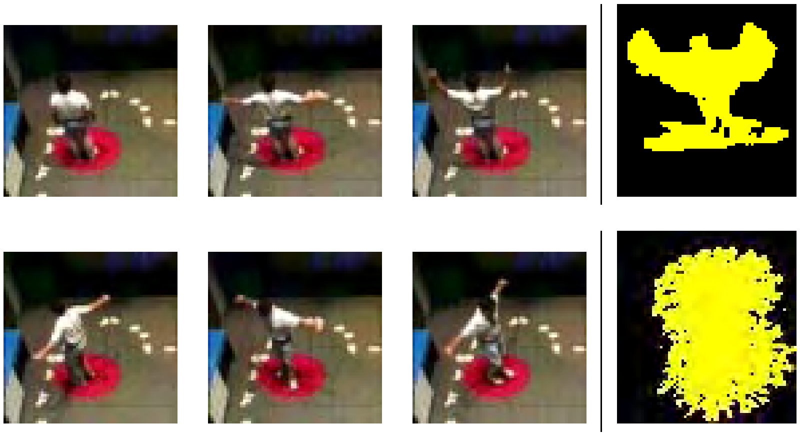


Figure 13: Two of the dance move actions are recognized using a motion template matching method [11]. The top left images show a person doing a flap move. The system detects the flap move by matching motion models (which have been computing using a database of example flap moves) to the motion template shown. Similarly, the bottom images show a person doing the spin move and the corresponding motion template. The top part of the blob is generated by the moving arms. The bottom part is generated by shadows from the arms. In the KidsRoom, shadows were incorporated into the models of the moves.

loop, much like those in the game industry and in commercial software like *Director* or *MAX*. The event loop continuously monitors the state of the room, checking all inputs as often as possible.

Dealing with real-time, physical interaction requires control structures more complex than those required in the typical keyboard-mouse situation, because actions take some amount of time during which the state of the actuating devices may change. The KidsRoom control structure partially handles those problems by using the notion of timers, and associating a timer with each event interval. Example uses of timers are ensuring that non-complimentary sounds do not play simultaneously, that background sounds appear continuous, and that narrations are spaced appropriately. The timing problems we encountered will be discussed in section 10.6.

8.2 Music and sound control

The KidsRoom has an original score written for this interactive installation. The music consists of 50 short MIDI segments, many of which can be concatenated to form musical phrases that gradually increase in complexity. The selection of musical segments, tempo, and volume is under computer control and is changed based upon the action in the room and the progression of the story. Computer control of the music is such that the control program can interrupt music abruptly or at the end of a musical phrase, depending upon the situation.

When the control program calls for a particular sound effect, the sound file is streamed to a process that adjusts the volume of the sound in the four speakers to localize the sound in a region of the room specified by the control program.

8.3 Lighting Control

The computer vision tracking and recognition algorithms require that the room be well-lit and that the lighting settings can be reliably set prior to each run. Consequently, special lighting is used only in transition segments during which time the vision algorithms are disabled. Even this modest use of lighting effects enhances the ambiance of the KidsRoom. The lighting is fully computer controlled using a MIDI light board. Some of our recent efforts in using automated vision systems in theater [25] use a multi-camera segmentation method that is invariant to lighting changes [19].

8.4 Animation Control

To capture the imaginative flavor of a story-book and to prevent the video effects from dominating the attention of the children, the KidsRoom uses layered, still-frame, cartoon-like animation sequences. The control program requests an animation, like "blue-monster-crouch" for a particular screen at a certain frame rate (usually about 2 frames per second), and several frames are streamed to the display. A benefit of such story-book animation is that we do not need to tightly synchronize the motion of the animated characters to that of the children, but, like a storybook, the still-frame cartoons can convey rich character activity.

8.5 Process Control

The KidsRoom control architecture is based on a client-server model. The control program is the client that communicates with ten servers to receive information about the state of the room and to control the output. The sensor servers are the object tracker server, the motion detector server, the two action recognition servers, and the scream detector server. The output servers are the directional sound server, the music server, the lights server, and the two display servers.

Communication is achieved using the RPC protocol. The server architecture has proven effective for allowing different individuals to work on different components of the system using the computer system most appropriate for the particular task. As noted by Coen [8], it is critical for any large distributed room control mechanism that individual components can be stopped and started without requiring a reboot of the entire system.

III. Evaluation and Analysis

In the remainder of this paper we evaluate the the KidsRoom with respect to our initial project goals and describe issues raised which would impact the construction of any similar environment.

9 Achieving project goals

We review the goals of Section 2 considering not only how well the goals were achieved but also the influence those goals had on the development of vision algorithms and the overall success of the project.

9.1 Real action, real objects

One of our primary goals was to construct an environment where action and attention was focused primarily in the room, not on the screens. We wanted a rich environment that would watch what the children do and respond in natural ways.

We believe the KidsRoom achieves this goal. Children are typically active when they are in the space, running from place to place, dancing, and acting out rowing and exploring fantasies. They interact with each other as much as they do with the virtual objects, and their exploration of the real space and the transformation of real objects (e.g. the bed) adds to the excitement of the play.

We were only partially successful in the use of real objects to enhance the experience. The only manipulated object in the KidsRoom is the bed; it is rolled around, jumped on, and hid behind, and it is a critical part of the narrative. However, two major obstacles — tracking and narrative control — prevented us from incorporating more objects into the room.

The KidsRoom tracking algorithm sets a limit of four people and one bed in the space because when more participants or large objects are present the space becomes visually cluttered, debilitating the tracking algorithm and interfering with perceptual routines. The second and perhaps more serious problem is that as more objects are added to a space, the behavior of the people in the space will become less predictable, because the number of ways in which objects may be used (or misused) increases. For the room to adequately model and respond to all of these scenarios it will require both especially clever story design and tremendous amounts of narration and control code. The more person-object interactions that the system fails to handle in a natural way, the less engaging and sentient the entire system feels.

To further achieve the goal of the keeping the action on the participants side of the screens, we designed the visual and audio feedback to only minimally focus the attention of the children. Typically, virtual reality systems use semi-realistic, three-dimensional rendered scenes and video as the primary form of system feedback. We decided, however, that in order to give the room a magical, theatrical feel and in order to keep the emphasis of the space in the room and not on the screens, images would have a two-dimensional story-book look and video would consist of simple, still-frame animations of those images. During much of the KidsRoom experience, the video screens are employed as mood-setting backdrops and not as the center of the participants' attention.

Audio is the main form of feedback in the room as sound does not require participants to focus on any particular part of the room. Children are free to listen to music, sound effects, and narration as they play, run about the space, and talk to one another. During the scenes where sound is the primary output mechanism, such as the bedroom and forest worlds, the children are focussed on their own activity in the space. Combining ambient sound effects with appropriate music can set a tone for the entire space. Audio feedback can be further enhanced by using spatial localization. Even with just four speakers, the KidsRoom monster growls sound like they are coming from the forest side of the room, and when the furniture speaks the sound originates from approximately the correct part of the room.

9.2 Remote visual sensing

In the KidsRoom there are no encumbrances on or requirements of people who enter the space except that they must enter one at a time. Further, occupants in the room (particularly young children) are typically unaware of how the room is sensing their behavior. There are no obvious sensors in the room embedded in any objects or the floor. The cameras are positioned high above the space, well out of the line-of sight and visible only if someone is looking for them. This enhances the magical nature of the room for all visitors, especially for children. They are not pushing buttons or sensors, they are just being themselves, and the room is responding.

9.3 Multiple, collaborating people

Another design goal was to create a system that could respond to the interaction of multiple people. Since self-consciousness seems to decrease as group size increases, the kind of role-playing encouraged by the KidsRoom is most natural and fun with a group. Also, when 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. For instance, during the rowing scene, children shout to one another about what to do, how fast to row, and where to row and play-act together as they hit virtual obstacles. Groups of friends and parent/child pairs have an especially good time.

9.4 Exploiting and controlling context

Given the difficulty of designing robust perceptual systems for recognizing action in complex environments, we strove to use narrative to provide context for the vision algorithms. Most of the vision algorithms depend upon the story to provide constraint. The boat rowing example described earlier typifies such a situation. It is currently well beyond the state of the art of computer vision to robustly recognize a group of closely situated people rowing a boat. In the KidsRoom, context makes it almost trivial.

Another example is the monster dance scene where the *story* was constructed to ensure that each camera has a non-occluded view of a child performing the dance moves. Potentially interfering children are cajoled by the monsters to stand in locations that do not interfere with the sensing. The advantage of an active system over that of a monitoring situation is the opportunity to not only know the context but to control it as well.

9.5 Presence, engagement, and imagination

The power of a compelling story-line cannot be overstated when constructing a space like the KidsRoom that integrates technology and narrative. The existence of a story seems to make people, particularly children, more likely to cooperate with the room than resist it and test its limits. In the KidsRoom a well-crafted story was used to make participants more likely to suspend disbelief and more curious and less apprehensive about what will happen next. The story ties the physical space, the participant's actions, and the different output media together into a coherent, rich, and therefore immersive experience.

Some existing work has studied the criteria that lead to the feeling of "immersion" or "presence" in virtual environments [28]. Here we just note that our system meets eight of the ten criteria commonly identified as important for creating a feeling of presence in a virtual space [29]. One of the two criteria the KidsRoom does not meet, "a similarity in visual appearance of the subjects and their representation in the virtual environment," does not apply to a system where people are interacting in the real world. Criteria for presence that are met include high-resolution information being presented to the appropriate senses, freedom from sensing devices, easily perceived effects of actions, an ability to change the environment, and "virtual" objects that respond spontaneously to the user. The remaining unmet criteria, that the system should adapt over time, is not met explicitly, but as discussed later the KidsRoom does allow the user to continuously and naturally control the pace of the experience.

An immersive space is most engaging when participants believe their actions are having an effect upon the environment by influencing the story. The KidsRoom uses computer vision to achieve this goal by making the room responsive to the position, movements, and actions of the children. Immediately upon their first interaction with the room the children realize that what they do makes a difference in how the room responds. This perceptual sensing enhances and energizes the narrative.

A goal that was critical to obtaining the presence of the KidsRoom was to *naturally* embed the perceptual constraints into the story-line. For example, in the monster dance scene, the vision systems require that there is only one child per rug and that all children are on some rug. One way to impose this constraint would have been to have the narrator say, "Only one kid per rug. Everyone must be on a rug." Instead, the *monsters* tell the children to stay on the rugs and that there can be only one per rug "so's we can see what's goin' on." That the monsters have some visual constraints seems perfectly natural and makes children less likely to feel restrained or to question why they need to engage in some particular behavior.

No matter how well an interactive story-line is designed, participants, especially children, will do the unexpected — especially when there are up to four of them interacting together. This unpredictable behavior can cause the perceptual system to perform poorly. Therefore, we designed the story so that such errors would not destroy the suspension of disbelief. When perceptual algorithms fail, the behavior of the entire room degrades gracefully.

One example of this principle in the KidsRoom is in the way the vision system provides feedback during the monster dance. If a child is ignoring the instructions of the characters and is too close to another child on a rug, the recognition of the movements of the child on the rug will be poor. Therefore, when actions are not recognized with high confidence, the monsters on the screen will animate, doing the low-confidence action, but the monster will not say anything. To the child, this just appears as though the monster is doing its own thing; it does not appear that the "monster" is any way confused. This choice was preferred over the possibility that the monster says "Great crouch" while the child is actually spinning.

Similarly, we tried to minimize the number of story segments that required a particular single action on the part of all participants. For instance, to get a particular piece of furniture to speak, only one child needed to be in its proximity; if all children needed to be close to it they might never discover how to make the furniture talk.

Finally, to give the KidsRoom narrative a cohesive, immersive feel, there are thematic threads that run throughout the story. For example, the careful observer will note that the stuffed animals on the walls in the children's bedroom, shown in Figure 4 are similar to the monsters that appear later in the story, shown in Figure 9. Some of the furniture characters in the first world have the same voices as the monsters in the monster world. The artwork has the same story-book motif in all four scenes, and several objects in the room on the shelves become part of the forest world backdrop during the transformation.

9.6 Children as subjects

Building a space for children was both wonderful and problematic. The positives include the tremendous enthusiasm with which the children participate, their willingness to play with peers they do not know, the delight they experience when being complimented by virtual characters, and their complete disregard of minor technical embarrassments that arose during development.

Children provided unique challenges as well. The behavior of children, particularly their group behavior, is difficult to predict. Further, children have short attention spans and often move about with explosive energy, leaving the longer playing narrations behind. Young children are small compared to adults, which can create problems when developing vision algorithms.

In balance, having children as the primary user group not only inspired us to think imaginatively but also, quite frankly, made the project all the more fun to construct.

10 Observations and failures

There were some issues that we failed to consider in the design phase of the KidsRoom that are important for developing other interactive, immersive spaces, particularly those for children. We present several in an effort to prevent others from repeating our mistakes.

10.1 Group vs. individual activity

The interaction in the KidsRoom changes significantly depending upon the number of people in the space. First, as mentioned previously, all system timings differ depending upon the number of people in the room. There is only a small window of time outside of which each unit of the experience becomes too short or too long – and the ideal timing changes based on the number of people around. Since automatically sensing when people are getting bored is well beyond our current perception capability, the KidsRoom uses an *ad hoc* procedure to adjust the duration of many activities depending upon the number of people in the space.

In general, the more children there are in the space, the more fast paced the room appears to be, because as soon as one child figures out the cause and effect relationship between some activity and response the other children will follow. A single child is more hesitant and therefore needs more time to explore before the room interjects. Also, a lone child often requires more intervention from the system to guide him or her through the experience.

A final consideration when developing for group activity is the importance of participants being able to understand cause and effect relationships. If too much is happening in the room and there is not a reasonable expectation within the child's mind of strong correlation between some action and a reasonable response, the child will not understand that he or she has caused the action to happen.

10.2 Exploratory vs. scripted narrative spaces

In our initial design of the KidsRoom, we planned to create a primarily exploratory space, modeled somewhat on popular non-linear computer games like Myst [7]. We designed and built prototypes for the first and second worlds using this model. In the first world, there was no talking furniture. Instead, when children walked near objects they made distinctive sounds: moving near the shelves with a mirror would make crashing sound, stepping on the rugs with animal pictures elicited the corresponding animal noises.

Our hope was that children would enter, figure out that they could make such sounds, and then explore the room, gradually creating a frenzy of sounds and activity. It didn't work. When we brought in some children for testing we found that they did not understand that they were causing the sounds – there was simply too much going on as each child explored independently. The same was true for some adults. Even children alone in the room had trouble identifying cause and effect relationships. We had also planned to develop an exploratory second world using forest sounds, forest images, and creepy, exploratory music. Again, testing proved the concept too weak.

The most significant problem was that the exploration "plot" did not encourage particular actions, nor did it cause people to act in a group fashion. Users did not share any common goals. Other authors have observed that exploratory, puzzle-solving spaces can sometimes make it difficult for adults to immerse themselves in an interactive world [14, 10]. When a *story* is added to the physical environment, a theatrical-like experience is created. Once the theatrical nature of the system is apparent, it is easier for people to imagine their roles and, if they are not too self-conscious, act them out. Furthermore, recognizing action is simpler in a story-based environment because the number of action possibilities at any moment is more constrained.

A *prima facie* criticism of a linear story-line is that the system loses its interactive nature. This is perhaps true for mundane interfaces such as mice and keyboards, as there are no dynamics to the actions performed. For a multi-person, room-sized environment, however, the interaction comes from making physical exertion, controlling the pace[11] of the adventure, recognizing that the room understands what is happening and is responding, and interacting with fellow users.

Finally, although the scenes and the "plot" are simple and linear, the actions within each scene are not. The system responds appropriately to user actions depending upon the context. Only the changes in context are linear.

10.3 Anticipating children's behavior

From testing with children, we learned that there are three aspects of children's behavior in the space that we had not adequately considered during narrative development.

---

[11] We intend a more psychologically loaded term than speed.

First, the story must take into account the children's "behavioral momentum." The KidsRoom is capable of making children exceedingly active. By the end of the first bedroom world when the furniture all start loudly chanting the magic word, the children are often running energetically around the room. Next the children end up on the bed, shout the magic word loudly, and the transformation occurs. The location is now the forest world and the children are instructed and expected to explore. However, the transformation typically calms them down and their tendency is often to remain on the bed. We found through testing that a fairly direct instruction (e.g. "Follow the path..."), sometimes repeated several times, is required to get them to start moving again. When designing for a space where physical action is the focus, behavioral momentum needs to be considered.

A related problem was the need for attention-grabbing cues. Particularly when kids are in a state of high physical activity, they almost never hear the first thing that the room says to them. Since we did not anticipate this problem, sometimes children missed important instructions. We modified the narrative so that it repeats some critical instructions more than once. Ideally, we should have built attention-grabbing narrative into the story-line for every critical narration.

Finally, children need to clearly understand the current task. The less certain they are of what to do, the more unpredictable their behavior becomes. Perhaps because they had never experienced a room such as this before, the children seem more inclined to wait for things to happen than to explore and try to make things happen. The children enjoyed the experience more once the system was modified so that there was always a clear task and when those tasks changed quickly.

10.4 Avoiding repetitiveness and hints

One way to break the suspension of disbelief of the experience is for the system to exactly repeat a single narration as it tries to encourage some behavior. Unfortunately, in a space like the KidsRoom built to encourage children to physically move around, instructions do need be repeated. For example, in the dance segment of the monster world, the control program continually checks if someone has stepped off their rug or if two people are on the same rug. If someone drifts off a rug more than once, narration is needed, but repeating a narration just played moments before imparts a mechanical sense to the responses and causes the entire experience to feel less alive.

One solution we developed was to use two different narrators. The main narrator has a deep, male voice and speaks in rhymes like a grandfather reading a story-book. The second narrator, with a soft, whispered, female voice, delivers "hints." The first time someone gets off a rug, the monsters will tell the person to get back on. After that, however, a voice whispers a hint, "Stay on your rug." This type of feedback is easily understood by room participants but does not break the flow of the story and primary narration. The delivery of hints by a different voice and typically from a different sound direction than the narrator made them perceptually salient, increasing their effectiveness. Because the hints were not long rhyming couplets it was easy to have multiple phrases to encourage a single behavior, reducing the repetition problem.

10.5 Perceptual limitations

Some perceptual-sensing difficulties and related issues follow.

10.5.1 Perceptually-based environmental constraints

A major challenge when designing the KidsRoom was to minimize the impact of our sensors and output modalities on the development of an interesting story environment. Some constraints are listed below.

- Vision algorithms generally require bright lighting, but large projection displays appear dim when placed in bright spaces.

- Large video screens displaying video violate the assumption of static background used by the vision algorithms. We were forced, therefore, to choose camera and rug positions so that people in the room would never appear in front of a screen in the image views – a serious limitation for a space like the KidsRoom or (even worse) the Cave [9]. Recent work motivated by this problem may alleviate this constraint [19, 12].

- Even in a space as large as 24 feet by 18 feet with a high ceiling, placing cameras so that the resulting views were non-occluded proved tricky. For example, although the overhead view provides a non-occluded view for tracking, we would prefer to find a tracking solution that does not require a 20 foot high ceiling. Further, due to viewpoint, occlusion, and story constraints, there is absolutely no flexibility in the positioning of the red and green rugs and their corresponding cameras.

- Every space-related decision required careful consideration of imaging requirements. For instance, rugs and carpet had to be short-haired, not shaggy, to prevent the background from changing as people moved around, and objects were painted in a flat paint to minimize specularity.

- The four speakers in the KidsRoom provide reasonable localization when listeners are near the center of the room. However, when a participant is near a loudspeaker playing a sound, that single speaker tends to dominate the positional percept and the spatial illusion breaks down. Sound localization is important if real objects in a space are to be given "personalities" using sound effects (e.g. as in the bedroom world), because the effect is destroyed if the sound is not perceived to come from the object.

- The KidsRoom has no mechanisms for understanding speech, only loudness. Adding unencumbering speech recognition capability to a space like the KidsRoom with loud music, loud sound effects, and loud children remains a significant challenge.

10.5.2 Object tracking difficulties

The KidsRoom tracking algorithm does an excellent job of keeping track of *where* people are but occasionally makes errors when keeping track of *who* people are [17]. In other words, the tracker sometimes swaps two people, thinking that one is the other, but in normal operation does not lose a person altogether.

The KidsRoom, therefore, was designed with the expectation that perfect tracking of identity might be problematic. The room uses information about where people are, but when referring to individuals it uses environmental indicators, not absolute labels. For instance, instead of saying, “Great job kid number one” it will say, “Great job, kid on the red rug.”

While there are some improvements that could be made to the tracking algorithm, perfect tracking of identity is unlikely. However, unlike conventional surveillance tasks, an immersive environment that must keep track of identity can manipulate people in its environment so that when it becomes uncertain of identity it can “bootstrap” itself automatically. For instance, a system controlling an environment with a telephone might physically call the room, ask to speak to a particular person and, when that person comes to the phone, reinitialize tracking. Integrating such “bootstrapping” devices into a narrative requires careful story development and will restrict the designer’s flexibility.

10.5.3 Monster world action recognition difficulties

All of the monster world action recognition strategies make assumptions about the viewing situation. First, images of the child cannot be occluded by other children. Careful camera and rug placement minimized this problem, although occasionally when large adults enter the space the system’s recognition performance will suffer slightly due to small occlusions. Second, since it is difficult to remove shadows accurately, shadows (and therefore light positions) were incorporated into the motion models. This turned out to yield a more robust representation but requires that the lighting setup doesn’t change between the training phase and run-time. Third, the motion-template recognition algorithms used in the KidsRoom are limited to recognizing actions of individual people, which prevents the interactive narrative from explicitly recognizing some multi-person action, such as people shaking hands.[12]

10.5.4 Event detection and non-cooperation

One problem we encountered when designing the KidsRoom was that “simple” events are strongly context-sensitive and memory-less. For example, our “in-a-group” detector will signal false continuously if one mischievous child refuses to cooperate with the remaining children. In this case, a more robust detector might ignore the renegade given that the child hasn’t been following the rules for a while. Similarly, if one child is scared and remains on the bed while other children explore the forest on the path, the “in-a-group” detector should ignore this child as well. We accommodated such possibilities not by fixing the detectors, which remains interesting future work in context-sensitive action recognition, but by ensuring that nowhere did the story stall endlessly if some requested or expected behavior was not observed.

10.6 Sensitivity to timing

Multiple people in a space increases the number of possible situations and responses that are required, thereby making narrative timing control difficult.

Our lack of any systematic approach to checking for inconsistent timings was most painfully apparent as we tested the nearly-completed system. Since advancing the story forward manually often creates timing problems, the only way to really test the room is to put people in it and run the narrative repeatedly. The problem is that the room provides a 10-12 minute experience, and thorough testing of every timing scenario is out of the question due to the large number of possible timing situations. Further, once the room is tuned, any small change to any timing-related code requires having several people around to interact in the space. Our only method of addressing this problem is to create modular story fragments such that the timers in one fragment do not affect those in another.

We note that an alternative mechanism to timers is to use the AI concept of planning to model the change in the states of the world [3]. Recently, Pinhanez, Mase, and Bobick [24] have proposed the use of *interval scripts*, where all the sensing and actuating activities are associated with temporal intervals, whose activation is determined by the result of a constraint satisfaction algorithm based on the input from the sensors and the past interaction. Such a representation may facilitate automatically checking for problematic temporally-related narrative control situations offline.

10.7 Communication model

The KidsRoom lacks a rich model of what information has been communicated to the users at any given time and how sequences of instructions can be presented to the users in a natural way. For example, in the river world, there are times when a room narration is presenting information to the children on the boat. In the middle of that narration, a child “jumps overboard” by getting off the bed. The system detects this event immediately, but has no way of promptly indicating this to the children. The solution is not as simple as cutting off the main narrator mid-phrase. First, abruptly cutting narrations destroys the sentient feel of the characters. Second, even if a narration can be naturally cutoff (e.g. using cue phrases like “oh!”), the system then needs a model of what partial information has been conveyed to the children and how to naturally pick up the narration when the overboard activity has ended. Significant time may have elapsed during the overboard activity, for example, which requires a modification of the original narration. Exactly how such a communication model should be represented and used for reasoning is an open research question.

10.8 Wasting sensing knowledge

Sometimes the system has the capability to detect some situation but no way of informing the participants. Given the impoverished sensing technology, no knowledge should be wasted – all information should be used to enhance the feeling of responsiveness.

For example, we encountered this problem when the children shout. Suppose the room asks for the kids to shout the magic word; the kids shout but not loudly. The room then responds with, “Try that shout one more time.” Initially we felt that children like to shout, so we would encourage them do it twice. The problem was that the room responded with ambiguous narration. Are they doing it again because it wasn’t good enough or for some other reason? Worse, if nobody shouted anything, the narration mildly suggested that the room actually heard a shout. As

---

[12] Group activity in the KidsRoom is always recognized using input from the person tracker, not using motion templates.

we improved the system, we added hints and new narration to try and ensure that when the system knows something (e.g. how loud a scream is) it lets the participants know that it knows. The effect is a room that feels more responsive.

Similarly, there are times when the system is dealing with uncooperative participants and, despite several attempts, has not elicited the desired activity. Another (shouting) example is when the narrator requests the children shout "Be quiet!" to the monsters. If after two requests the children do not shout, the story continues, ignoring their lack of cooperation. However, it is important to explicitly acknowledge that the system understands that something is wrong but is ignoring the problem. In this example, the narrator says *"Well, I can't say* that *was a very loud shout, but perhaps the monsters will figure it out."*

10.9 Perceptual expectation

Given the technological limitations of unencumbered sensing, the interaction must be designed as to not establish any perceptual expectations on the part of the user that cannot be satisfied. For instance, use of some speech recognition in the KidsRoom might prove problematic. If a child or adult sees that the room can respond to one sentence, the expectation may be established that the characters can understand *any* utterance. Any immersive environment which encourages or requires people to test the limits of the perception system is more likely to feel more mechanical than natural.

The KidsRoom is not entirely immune to this problem, but we believe we have minimized the "responsiveness testing" that people do by making the system flexible to the type of input it receives (e.g. in the boat scene most any large body movement will be interpreted as rowing) and by having characters in the story essentially teach the participants what they can and cannot recognize (i.e. the allowed dance moves in the monster land).

11 Summary and contributions

The KidsRoom went from white-board sketches to a fullly-operational installation in eight weeks. This paper has described the story, technology and design decisions that went into building the system. We believe the KidsRoom is the first perceptually-based, multi-person, fully-automated interactive, narrative playspace ever constructed, and the experience we acquired designing and building the space has allowed us to identify some major questions and to propose a few solutions that should simplify construction of complex spaces in the future.

We believe the KidsRoom provides several fundamental contributions. First, unlike most previous interactive systems, the KidsRoom does not require the user to wear special clothing, gloves or vests, does not require embedding sensors in objects, and has been explicitly designed to allow for multiple simultaneous users.

Second, it demonstrates that non-encumbering sensors can be used for the measurement and *recognition of individual and group action* in a rich, interactive, story-based experience. Relatively simply perceptual routines integrated carefully into a strong story context are adequate for recognizing many types of actions in interactive spaces.

Finally, we believe the KidsRoom is a unique and fun children's environment that merges the mystery and fantasy of children's stories and theater with the spontaneity and collaborative nature of real-world physical play.

Credits and Acknowledgments

The KidsRoom owes much of its success to the fact that its concept and design was developed, collaboratively, by all of the authors. Aaron Bobick acted as advisor and producer. Stephen Intille (interaction control; production issues) and Jim Davis (vision systems) were the chief architects of the system. Freedom Baird, with help from Arjan Schütte, wrote the original script, handled most narrative recording, and provided video documentation. Claudio Pinhanez was in charge of music and light design and control. Lee Campbell worked on sound control and related issues. Yuri Ivanov and Andrew Wilson wrote the animation control code. Composer John Klein wrote the original KidsRoom score, and artist Alex Weissman created the computer illustrations.

The authors would like to thank: Glorianna Davenport for her support; Ed McCluney, John UnderKoffler, Laurie Ward, and Michelle Kemper for character narrations; Hugo Barra for work on MIDI light control; Greg Tucker; and Laurie Ward for administrative support. We also owe thanks to OpCode and Mark of the Unicorn for software donations. This work was funded by the MIT Media Laboratory's Digital Life Consortium.

References

[1] R.T. Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385, August 1997.

[2] J.M. Barrie. *Peter Pan*. EP Dutton, 1988.

[3] Joseph Bates, A. Bryan Loyall, and W. Scott Reilly. An architecture for action, emotion, and social behavior. In *Proc. of the Fourth European Workshop on Modeling Autonomous Agents in a Multi-Agent World*, S. Martino al Cimino, Italy, July 1992.

[4] B.B. Bederson and A. Druin. Computer augmented environments: New places to learn, work, and play. In *Advances in Human-Computer Interaction*, volume 5, chapter 2. Ablex Pub. Crop., Norwood, NJ, 1995.

[5] Aaron F. Bobick. Movement, activity, and action: the role of knowledge in the perception of motion. *Phil. Trans. Royal Society London B*, 352:1257–1265, 1997.

[6] R.A. Bolt. *The Human Interface*. Wadsworth, Inc., Belmont, California, 1984.

[7] Broderbund Software. Myst. An interactive CD-ROM, 1994.

[8] M.H. Coen. Building brains for rooms: designing distributed software agents. In *Proc. of the Conf. on Innovative Applications of Artificial Intelligence*, pages 971–977. AAAI Press, 1997.

[9] C. Cruz-Neira, D.J. Sandin, and T.A. DeFanti. Surround-screen projection-based virtual reality: The design and implementation of the CAVE. In *Proc. of SIGGRAPH Computer Graphics Conference*, pages 135–142. ACM SIGGRAPH, August 1993.

Figure 14: Kids of all ages have "played" in the KidsRoom. Children have been willing to wait in line up to one hour to experience the room a second and third time.

[10] G. Davenport and G. Friedlander. Interactive transformational environments: Wheel of life. In E. Barrett and M. Redmond, editors, *Contextual media: multimedia and interpretation*, chapter 1, pages 1–25. MIT Press, Cambridge, MA USA, 1995.

[11] J.W. Davis and A.F. Bobick. The representation and recognition of action using temporal templates. In *Proc. Computer Vision and Pattern Recognition*, pages 928–934. IEEE Computer Society Press, June 1997.

[12] J.W. Davis and A.F. Bobick. A robust human-silhouette extraction technique for interactive virtual environments. In *Lecture Notes in Artificial Intelligence*, volume 1537, pages 12–25. Springer-Verlag, 1998.

[13] A. Druin. Noobie: the animal design playstation. In *Proc. of Human Factors in Computing Systems (CHI)*, volume 20, pages 45–53, 1988.

[14] A. Druin and K. Perlin. Immersive environments: a physical approach to the computer interface. In *Proc. of Human Factors in Computing Systems (CHI)*, pages 325–326, April 1994.

[15] J. Glos and M. Umaschi. Once upon an object...: computationally-augmented toys for storytelling. In *Proc. of the Int'l Conf. on Computational Intelligence and Multimedia Applications (ICCIMA)*, pages 245–249, February 1997.

[16] M. Hu. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, IT-8(2), 1962.

[17] S.S. Intille, J.W. Davis, and A.F. Bobick. Real-time closed-world tracking. In *Proceedings of the IEEE Computer Society Converence on Computer Vision and Pattern Recognition*, pages 697–703, Los Alamitos, CA, June 1997. IEEE Computer Society, IEEE Computer Society Press.

[18] H. Ishii and B. Ullmer. Tangible bits: towards seamless interfaces between people, bits, and atoms. In *Proc. of Human Factors in Computing Systems (CHI)*, pages 234–241, March 1997.

[19] Y. Ivanov, A.F. Bobick, and J. Liu. Fast lighting independent background subtraction. In *IEEE Workshop on Visual Surveillance – VS'98*, pages 49–55, January 1998. Also appears as MIT Media Lab Perceptual Computing Group TR#437.

[20] M.W. Krueger. Environmental technology: Making the real world virtual. In *Communications of the ACM*, volume 36, pages 36–37, 1993.

[21] Margot Lovejoy. *Postmodern Currents: Art and Artists in the Age of Electronic Media*. Ann Arbour, London, England, 1989.

[22] P. Maes, A. Pentland, B. Blumberg, T. Darrell, J. Brown, and J. Yoon. ALIVE: Artificial life interactive video environment. *Intercommunication*, 7:48–49, 1994.

[23] A. Pentland. Smart rooms. *Scientific American*, 274(4):68–76, April 1996.

[24] Claudio S. Pinhanez, Kenji Mase, and Aaron F. Bobick. Interval scripts: a design paradigm for story-based interactive systems. In *Proc. of Human Factors*

in Computing Systems (CHI), pages 287–294, March 1997.

[25] C.S. Pinhanez and A.F. Bobick. It/I: Theatre with an automatic and reactive computer graphics character. In *Proc. of SIGGRAPH'98 Sketches*, July 1998.

[26] Frank Popper. *Art of the Electronic Age*. Thames and Hudson, London, England, 1993.

[27] M. Sendak. *Where the Wild Things Are*. Harpercollins Juvenile Books, 1963.

[28] T.B. Sheridan. Musings on telepresence and virtual presence. *Presence: Teleoperators and Virtual Environments*, 1(1):120–125, 1992.

[29] M. Slater and M. Usoh. Presence in immersive virtual environments. In *IEEE Virtual Reality Annual International Symposium*, pages 90–96, 1993.

[30] Christa Sommerer and Laurent Mignonneau. Art as a living system. *Leonardo*, 30(5), October 1997.

[31] Mark C. Torrance. Advances in human-computer interaction: the intelligent room. In *Working notes of the CHI 95 research symposium*, May 1995.

[32] Naoko Tosa, Hideki Hashimoto, Kaoru Sezaki, Yasuharu Kunii, Toyotoshi Yamada, Kotaro Sabe, Ryosuke Nishino, Hiroshi Harashima, and Fumio Harashima. Network-based neuro-baby with robotic hand. In *Proc. of IJCAI'95 Workshop on Entertainment and AI/Alife*, August 1995.

[33] M. Umaschi. Soft toys with computer hearts: building personal storytelling environments. In *CHI Extended Abstracts*, pages 20–21. ACM, 1997.

[34] Walt Disney Productions. Bedknobs and broomsticks. Movie, 1971.

[35] M. Weiser. Some computer science issues in ubiquitous computing. In *Communications of the ACM*, volume 36, pages 74–84, 1993.

[36] 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.