M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 398
November, 1996
Revised August 1997

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

Aaron Bobick, Stephen Intille, Jim Davis
Freedom Baird, Claudio Pinhanez
Lee Campbell, Yuri 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 perceptual, 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]

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; a living room might warn parents when a child is engaging in some potentially dangerous activity; finally, a child's bedroom might stimulate a child's imagination by using images and sound to transform itself into a fantasy world.

In the spirit of several popular children's adventure books, we have built the KidsRoom, a fully-automated, interactive, narrative playspace for children. Two of the bedroom walls resemble the real walls in a child's room, complete with real furniture. The other two walls are large back-projected video screens used to "transform" the room environment. Four speakers project sound effects, music, and narration into the space. Three video cameras observe the space and are used by computer vision algorithms to track the positions of the children and recognize their actions. Computer-controlled lighting provides illumination, and a microphone is used to detect the volume of children yelling. The room is controlled by six computers. Three are dedicated to vision processing, and the others handle narrative control, video, lighting, and sound.

The KidsRoom experience lasts 10-12 minutes, depending upon how the participants act in the room, and it was designed primarily for children ages 6-8 years old. Throughout the story, children interact with objects in the room, with one another, and with virtual creatures projected onto the walls. The actions of the children taking place in the real physical space drive the narrative action forward. The room reacts to the activity of children, and the children are aware that the room is responsive. Simultaneously, the narrative context of the story is used to control the complexity of context-dependent action recognition algorithms.

[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.

This paper describes the KidsRoom, the technology that makes it work, and some observations that may be of interest to other designers of perceptually-based interactive, immersive systems. Section 2 reviews some previous work relevant to constructing interactive, immersive environments. Section 3 discusses the difference between sensing movement and understanding action and describes what we believe is the primary contribution of this work. Section 4 outlines the major design goals of our system and why we chose to construct a children's playspace. Section 5 describes the physical layout of the KidsRoom. Section 6 briefly outlines the first three segments of the KidsRoom story and includes a detailed, annotated example of the last scene. The computer vision algorithms that provide sensory input to the KidsRoom are discussed in section 7, and the story control mechanisms and related issues are described in section 8. Finally, we discuss some problems we encountered and some other observations in section 9. Section 10 is a short summary.

2 Interaction in augmented environments

The KidsRoom was designed to provide an immersive, narrative environment for children in an augmented physical space as opposed to a virtual space. 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. In such "augmented reality" systems, feedback is normally provided by using head-mounted displays that superimpose graphical displays on real imagery[1]. Most common are systems where users observe "virtual" action in a virtual space. The Cave, consisting of four video walls and a head-mounted sensor is one popular system of this form[8].

We are interested in systems of the second form, where the computer interface becomes unobtrusively embedded in the physical world itself . Users interact in a space as they normally would, performing and observing physical body actions and receiving feedback naturally integrated into the physical and virtual environments. Such systems have been called (among other things) "augmented environments," "immersive environments", "intelligent rooms,"[31] "smart rooms,"[22] and "interactive spaces."

One early example of augmenting a physical space was the "Media Room" project of Bolt and Negroponte[6]. They summarize their goal as follows:

> "The human interface with computers is the physical, sensory, and intellectual space that lies between computers and ourselves. Like any place this space can be unfamiliar, cold and unwelcoming. But it can also be like some other places, those we know and love, those that are familiar, comfortable, warm, and most importantly, personal."

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.

Research on physical, interactive spaces that do not require cumbersome body sensing devices began with Krueger's Videoplace system[18]. Krueger designed installations that explored many different modes of interaction, including one where the user interacted with his or her own silhouette on video screens. In another installation, the user's image was affected by the hand of another user. A third installation included a simple form of computer-controlled character which would play with the contour of the user's image.

Krueger was exploring interfaces that respond to large body gestures. Others began to investigate how computer interaction changed when computer interaction was mediated through real, physical objects. Druin, for instance, constructed a stuffed-animal doll called Noobie[11]. Children could interact with Noobie by squeezing Noobie's limbs and watching a display embedded in Noobie's belly. The response of the computer depended upon how the children interacted with a real object that sat in their surrounding physical space. 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 they were already comfortable with. The children were not encumbered with sensors in any way (also see work by Umaschi and Glos[33, 13]). A similar project for children is the Lego-Logo research at the MIT Media Laboratory. Lego devices have been constructed that are motorized and computer-controlled so that the computer can be brought into the world of play of children[28].

Druin and Perlin set out to construct immersive physical environments for adults that responded to movement and touch in real physical spaces[12]. 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.

Narrative in interactive physical spaces was further explored by Davenport and Friedlander [9]. They wanted people to "feel as though they were walking through a computer monitor into a magic landscape." They constructed a four-world, semi-automated 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 experi-

ments are available[19, 26]. Noteworthy installations include Jeffrey Shaw's *The Legible City*, where the user physically rides a bicycle through a virtual city of poetry, and 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. The ATR–MIC laboratory in Japan produced some of the most technology-intensive works. Artists Christa Sommerer, Laurent Mignonneau, and Naoko Tosa integrated computer vision, computer graphics, genetic algorithms, and emotion and speech recognition techniques[30, 32].

Several groups have explored interactive spaces that are naturally integrated into office environments. At Xerox PARC active badge technology[34] has been used so that computers can track a worker's whereabouts and bring the computing environment to the user instead of forcing the user to go to the computing environment. Interfaces modeled on the paradigm of post-it notes (PARCTab), paper (MPad), and whiteboards (LiveBoard) have been constructed with the presumption that people like the way they communicate now. Instead of forcing people to change, it is possible to computationally augment devices they already understand giving people and devices more flexibility and capability[35]. Another project at PARC has investigated rooms that respond to occupants by automatically adjusting the physical environment by manipulating heat, light, and power[4].

Wellner has explored related ideas by designing a desk that uses image processing to track a person's finger and recognize text on papers tracked on a desk[36]. The system responds by integrating the real and virtual in the real physical space. For example, the PaperPaint program allows someone to make a real drawing and then virtually copy and paste the drawing on the desk. More recently, Ishii and Ulmer have designed the MetaDesk[16]. The MetaDesk user can manipulate supplementary virtual displays by naturally manipulating real physical objects like blocks and flashlights. The goal of both of these projects is a seamless interaction between real and virtual spaces .

The ALIVE project at the MIT Media Laboratory 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[20]. 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" project at the MIT AI Laboratory exists to investigate the control mechanisms and sensing technologies required to design reactive spaces[31]. The system consists of several cameras and two large screens in a small room. A single 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[7]. The goal of the project is to remove the computer from the human-computer interface.

Bederson and Druin believe, as do we, that the best immersive physical environments will have multimodal inputs and outputs[4]. Increasing computational speed is now making it possible to explore domains like immersive office environments, living spaces, and theater performances[24]. The two major obstacles to building fully-automated reactive spaces are (1) finding practical, computationally-feasible sensing modalities that can 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.

3 Primary contribution

The primary contribution of the KidsRoom is its demonstration that non-encumbering sensors can be used for the recognition of individual and group action in a complex interactive narrative. Relatively simply perceptual routines integrated carefully into a strong story context are adequate for constructing some types of rich interactive and immersive spaces. In this paper we fully document our design process, problems we encountered, and solutions that we developed with the hope that our observations will help future interactive space developers anticipate problems and improve upon our successes.

The systems described in the previous section all use sensors to measure the body positions of the system participants. Systems that do so without requiring the user to wear special clothing, gloves or vests and without embedding sensors in objects typically only handle one person at a time (e.g. Alive[20]). The KidsRoom uses computer vision to measure positional and motion information of multiple, interacting objects, which most previous vision-based virtual environments have not done.

The KidsRoom begins to move beyond just measurement of position towards recognition of action using measurement *and* context. 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[37]. Although the mechanisms used are simple, the KidsRoom combines the sensor outputs with contextual information derived from the story to to recognizing over a dozen simple individual and group actions in specific contexts. Those recognized actions are then used to further drive the story and control a 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 (see Bobick's taxonomy of movement, activity, and action for an example[5]). Many actions of interest require that contextual knowledge be used for recognition in addition to sensor 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 the same measurements may correspond to different actions. Stronger contextual constraints are required to extract action labels from perceptual measurements. The KidsRoom's environment is rich enough to begin to explore some context-sensitive recognition tasks.

Most of the actions recognized by the KidsRoom are of the simplest type, where motion measurement provides the majority of the information used to identify the action. However, context is used extensively when checking for necessary conditions that determine if particular types of action recognition methods are valid. Further, some highly context-sensitive actions (e.g. "staying in a group") are recognized, albeit in a simple manner. The KidsRoom, therefore, marks a first step towards using context for action recognition in a fully-automated, perceptually-based immersive environment.

4 Design goals

We had several goals for our system which are introduced in this section and revisited later in the paper. Our first goal was that, in the spirit of most of the work reviewed above, the system recognize action 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 that people are comfortable with.

One way to make it easier for the focus of activity to remain in the physical space is to not encumber a room's occupants by requiring them to wear sensors, head-mounted displays, earphones, microphones, or specially-colored clothing. As our second goal, we were committed to using computer vision tracking and action recognition software to allow any person to easily enter and exit the room at any time naturally without troublesome sensor requirements.

Why not use sensors embedded in objects in the space? For example, we might have used touch sensors on furniture and in the floor. First, these types of sensors do not allow the system to observe and interpret the physical interaction between the people in the space. Floor sensors and touch sensors cannot be used to discriminate between many different body gestures, and these are often the same actions that are of the greatest interest to a system that is trying to interpret and respond to human interaction. Second, we believe that while sensor integration may be practical for simplified interactive spaces, physically wiring every object in a complex space with several different devices will ultimately prove to be impractical and expensive. A single camera with the appropriate visual processing software may be able to do the job of thousands of distributed sensors. Finally, even if every device in a room is wired with sensing devices, the difficult problem of *understanding what is happening* in the space has not been solved. Some program has to combine the perceptual inputs and then, much like computer vision programs try to do, make sense of the world. For these reasons we chose to user visual sensing.

A third goal was to design a space that could handle the interaction of multiple people and objects. Previous work in computer vision has dealt primarily with environments containing only one and sometimes two people. Similarly, most previous work on fully-automated interactive spaces could not handle multi-person interaction. However, immersive spaces for living environments or office environments will require such capability. We wanted to construct a system that enriched and stimulated the spontaneous, real-world intercommunication and interaction that groups of people enjoy. Similarly, we wanted to design an immersive system where people could manipulate real, everyday objects that were somehow tied to a virtual environment. Our hope was to seamlessly combine the real and the virtual interaction.

As explained above, we wanted non-encumbering computer-vision action recognition algorithms to provide sensory input to the immersive environment's control architecture. However,

Since vision-based perceptual systems for multi-person, room-like environments are challenging to construct given our restriction of non-encumbering sensing, an additional design goal was to build a system where the narrative provides contextual constraints used to make the vision-based perceptual routines more robust.[2] At the same time, we believed the visually-recognized activity could be used to convincingly guide the narrative forward and give participants a feeling that they room was responding to their actions.

Finally, we strove to create a room environment that was truly "immersive." We worked to avoid situations were people had no idea how to proceed and therefore had to ask for outside help. The experience encourages people to interact with everyday, familiar objects during imaginative, immersive role-playing. Further, the system is designed so that when unexpected situations occur that are not understood by the perception algorithms the system proceeds in a way that does not break the immersive feel.

After establishing these goals and carefully considering several immersive environment scenarios, we

[2] An example of a non-narrative space would be a kitchen that watches what people do and responds by offering advice. This task provides weaker constraints for vision algorithms than a situation where people are engaged in some type of linear narrative. The narrative will focus and constrain each person's behavior somewhat so that future actions are limited to a small number of predictable possibilities. Vision algorithms can then use this expected context to make simplify difficult sensing situations. In a kitchen however, complex and poorly understood models of human behavior are required to predict what a person might do at any moment to provide similarly powerful constraints for vision algorithms.

converged on the idea of constructing a children's playspace. We observed that children already like to play with each other in real spaces engaging in rigorous physical activity, and that they would be more likely than adults to be motivated by supplementary imagery, sound and lighting. We also felt that children would be less prone to feel self-conscious once the interaction began and that children would enjoy making the types of large body gestures we thought our system could recognize. Further, we thought that children would be less-sensitive to small glitches in timing, animations, and recognition and more focused on having fun in the space than on figuring out how it works or how to break it.

Davenport and Friedlander, and Druin and Perlin both observed that adults visiting their interactive installations sometimes had difficulty immersing themselves in the narrative. Davenport and Friedlander found that it was difficult to make some people comfortable so they would "play along" without getting embarrassed[9]. 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!" Once they are in a comfortable environment, children are less-likely than adults to be self-conscious or to be hesitant to move and manipulate objects.

Building a space for children is not without its problems, however. 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. Young children are small compared to adults, which can create problems for vision algorithms. Finally, critical incremental testing during development could only be performed occasionally, when children could be brought to our lab.

5 The Playspace

The KidsRoom re-creates a child's bedroom. The space is 24 x 18 feet with a wire-grid ceiling 27 feet high. Two of the bedroom walls resemble the real walls in a child's room, complete with real furniture, posters, and window frames. 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 directional 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. The other furniture, which is not explicitly tracked by the computer system, is fastened to the ground and nailed shut.[3] Four colored rugs with animal drawings and simulated stone markers on the floor are used as reference points during the narrative. 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 for action recognition when children are standing on the green and red rugs, respectively. Finally, camera four provides a nice view of most of the room and the two screens for spectators outside the space. In addition to the visual input, a single microphone is in the space that is used to detect the loudness of screams during two parts of the KidsRoom narrative.

The room has five types of output for motivating participants: video, music, narration, sound effects, and lighting. Video is projected on the two walls. Music is used to enliven the space, set the mood, and dynamically alter dramatic tension throughout the story. Narration and other sound effects are directional, appearing to come from particular regions of the room. Further, some effects, like monster growling and boat crashes are particularly loud and can vibrate the floor. As we discuss later, lighting must remain constant when the vision algorithms are operating, however, since 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 at one room position, motion energy detection, sending MIDI music commands to the music computer, and amplitude audio detection. A third SGI Indy workstation is used for action recognition at a second room position. Two Digital AlphaStations are used for displaying still-frame animations, one per screen. One of the AlphaStations also runs the room's control process. Finally, one Macintosh is used for running Studio Pro MIDI software.[4]

6 The Story

The KidsRoom guides children through an interactive, imaginative adventure. Inspired by three famous children's stories in which children's rooms are transformed into magical environments[2, 27, 29], the story begins in a child's bedroom and progresses through three other worlds. We will describe the last world in

[3] Movement of furniture other than the bed is not modeled by the system.

[4] We have used the hardware we had available. However, it might be possible to replace the SGI workstations with less costly computers (e.g. PCs) using special video hardware, and it is certainly possible to replace the AlphaStations and Macintosh with PCs for displaying images and playing MIDI music.

Other essential equipment used by the room includes the 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 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. Six computers used to control the fully-automated room and additional audio and video equipment is located just outside one corner of the room. 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.

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. 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.

detail, to give the reader a feel for the story, it's characters, and the interactive responsiveness of the entire system.

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 attends to this region, checking for people entering and exiting the space. Curious music plays softly, and the projection walls display scenes from a bedroom, as shown in Figure 4. The children are told only to "ask the furniture for the magic word." The magic word is used to trigger the start of an adventure. When at least one child approaches some piece

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.

Figure 4: (a,b) When participants first enter the space, the projection walls display scenes from a bedroom. (c) Children are told only to "ask the furniture for the magic word."

of furniture (e.g. the blue desk or the green frog rug), each of which has a distinct personality, the furniture speaks. For example, a chest says, "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 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. If the children get confused and don't 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 children. Eventually, one piece of furniture knows the magic word, which alternates between several possibilities each time the system is run, and tells the children. Then all of the furniture start chanting the word loudly. A mother's voice breaks in, silencing the furniture voices, and tells 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.

6.2 The forest world

After the kids yell the magic word loudly, the room darkens and the transformation occurs. Images on the projection walls gradually fade to images of a cartoon fantasy forest land. Colored, flashing lights combined with mysterious music play during the transformation, as shown in Figure 5a. Simultaneously, a deep-voiced narrator (the room) says, "Welcome to the KidsRoom. It's not what it seems. What you might see here are things dreamt in your dreams."

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. The children are told to stay in a group and to stay on that path. The system responds with encouraging

(a) (b)

Figure 5: (a) Lighting effects are used to mark special transitions in the story. These images were taken as the room transforms from a bedroom to a forest world. The lights go out and transitional music plays. A deep-voiced narrator speaks as colored lights flash and the screens gradually fade from bedroom to forest. "Welcome to the KidsRoom. It's not what it seems. What you might see here are things dreamt in your dreams." (b) During the forest world, children must hide behind the bed to stop the loud growling of distant monsters.

narration if they don't do so. As they traverse the path (moving around the room several times), monsters are heard growling from afar. The children are told, "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 don't hide, the loud growling continues and narration encourages the kids to get behind the bed. After a short walk, the children reach the river world.

6.3 The river world

Images of a river fade onto the two screens. One view shows the river progressing forward (see Figure 6) and the other view shows the 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 yells, "Passenger overboard" 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 6. The children must engage in collaborative rowing (making rowing motions on the correct side of the bed) to avoid the obstacles. When they do, a female voice softly whispers hints like "Nice job." When they don't, they hear a loud crashing sound, often motivating the kids to physically play-act a crash as in Figure 6b, and they receive some whispered hints about what to do.

Eventually, the children reach the monster world. They are told 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 it is being pushed onto the sand, and 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. When they do, the monster land appears.

6.4 The monster world

This section will describe the monster world in more complete narrative detail. As the world begins, forest images are displayed on the two screens. Creepy forest music plays in the background, and jungle sounds like twigs cracking and owls are can be heard. The room speaks in a comforting, deep, somewhat mischievous voice.

> Room: "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 7a-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.

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

If the kids still don't scream, the room says, "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.

> Room: "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, the room continues, "Always face the monster that's closest to you, and get ready to do what the monsters do."

The vision algorithm needs to have just one child on each rug to get a non-occluded view of each participant. Therefore, if the system detects multiple people on a rug, the monster characters in the story respond.

> Monster: "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.

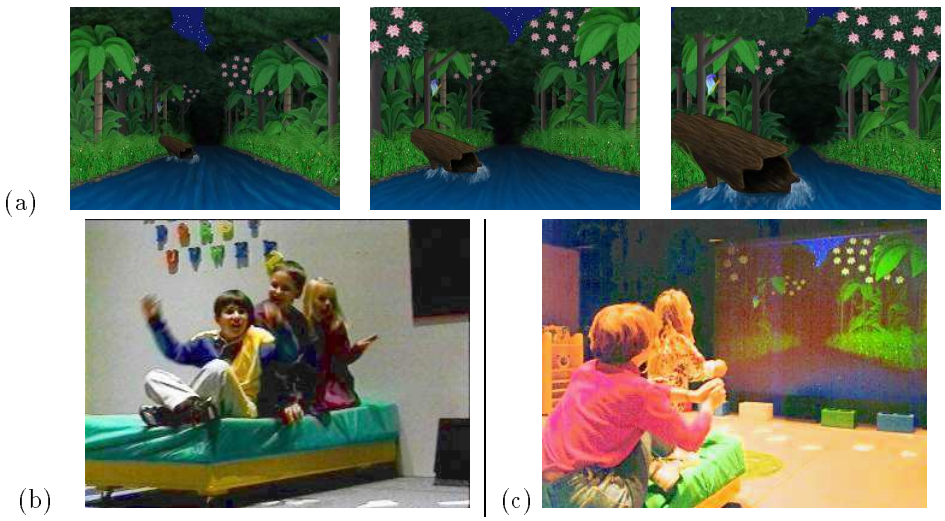

Figure 6: (a) The speed of the boat is controlled by how vigorously people on the bed are rowing. If everyone stops making motions, the boat imagery will stop moving forward. Obstacles such as the log in this series of images approach the boat. They can be avoided by rowing strongly on the appropriate 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.

Figure 7: (a) A child dancing with a monster. (b) Children spinning during the monster dance.

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

When each rug has a single child on it, the monsters continue. They teach the children four dance moves. Energetic music and forest sounds continue to play in the background.

> 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 8 does the move, and then says, "Your turn." The system watches the kids on the yellow and green rugs.[5] If a kid is detected to have done the crouch move, the monsters respond.

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

The monsters continue, teaching the children three other moves: the Y move, the flapping move, and the spinning move. The monsters only respond when the system detects the moves. Once the children know the moves, the music turns more energetic.

> Room: "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."

[5] 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, since 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 that case children never expressed disappointment that the system was only responding to the children on the red and green rugs.

Figure 8: 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 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, and if the system is especially certain it has recognized a move, the monsters will offer spoken confirmation.

Now, if the kid on the red rug starts spinning, the blue monster spins and says, "Super spin!" If the kid on the green rug does a crouch, then the brown monster does a crouch saying, "Great crouch!" When a kid on the green rug does some flaps the brown monster says, "Awesome flaps." and does some flaps. If both kids stop doing recognizable moves, a whispered hint can be heard, "Try a flap, spin, Y, or crouch."

This dance phase continues for a few minutes (see Figure 7). The music grows louder and faster and the room says, "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 their own "monster moves."

The children run around the room and do new dance moves. Suddenly, the mother's voice is heard again, and all music and sound abruptly ends. She says, "I told you kids to go to bed, and I mean business."

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

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 lights drop down and colored transition lighting is used. The transformational music plays, as the monsters say goodbye, "Take it easy! Bye bye!" The forest fades back to the room as shown in Figure 9, and as the lights slowly come up, the room says, "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 3, in the KidsRoom we both *measure* the movement of multiple, interacting people and then use that data and contextual information to *recognize* action using vision-based perceptual algorithms. This section describes the perceptual recognition algorithms used by the KidsRoom and some difficulties we encountered with each method.

7.1 Object Tracking

Most useful or entertaining 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. For example, the blue desk will speak only when a person is near it. The positional information is also used to keep track of whether people are in a group and whether they are moving or not. Finally, position information is used to create known contexts for other vision processes by ensuring that people are in expected regions of the room.

7.1.1 Method

The KidsRoom tracking algorithm uses the top-down camera view of the space, which minimizes object occlusion. Further, lighting is assumed to remain constant during the time that the tracker is running. We have tried to light the space carefully from above so that brightness remains constant from one side of the room to the other, 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. It is important for the algorithm to keep track of how many people are in the room, therefore, which is achieved by having everyone in the room enter and exit through a "door" region. The context-sensitive, non-rigid object tracking method is fully described and tested elsewhere[15]. The bed is tracked as well and can be rolled around the room.

Figure 9: Images on the projection screens are used to "transform" the bedroom into the fantasy forest worlds and visa versa. 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.

Figure 10 shows the image view used for tracking and the output of the tracking system. Each person is represented by a single point, the centroid of his or her blob. The box in the lower left corner represents the "door" region of the room, where people can enter and exit at any time. The tracking algorithm is not limited to four people, but the KidsRoom narrative was designed for a maximum of four participants.

7.1.2 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[15]. In other words, the tracker sometimes swaps two people, thinking that one is the other, but it will rarely lose a person altogether.[6]

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[15], flawless tracking of identity for several minutes is unlikely. However, unlike conventional surveillance tasks, immersive environments 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.

[6]The few occasions where the tracker does lose track of where there is a person typically result from violations of expectations built into the system. One expectation is that a six-year-old child will be at least a certain size in the image. Therefore, when a particularly young child (i.e. younger than 6) is in the space the child can sometimes appear so small that the system treats the child's blob as noise.

7.2 Movement detection

In addition to tracking object positions, a useful perceptual tool for an immersive environment is the ability to recognize action by simply recognizing movement in a strongly constrained context. In the river world of the KidsRoom, 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.

Knowing a specific context setup by the story makes it possible to detect a relatively complex action using a simple motion detection algorithm. The rowing detection algorithm requires that everyone is "inside the boat," or all on the bed. The narrative encourages participants to setup this context, and a simple visual algorithm is used to confirm that the context is achieved. To do this, the algorithm uses the same top-down, background-subtracted image used by the tracker. Since the size of the bed is known, the algorithm waits until the number of pixels in the large blob created by the bed and the people is approximately equal to the known size of the bed. The story is used to encourage participants to be inside the bed's boundary (e.g "Tuck your hands and feet right in, the hungry sharks are eager to sin.") When the blob size is about right, everyone is assumed "in the boat" and the bed orientation (i.e. left and right side of bed) 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, both, and right sides of the bed.

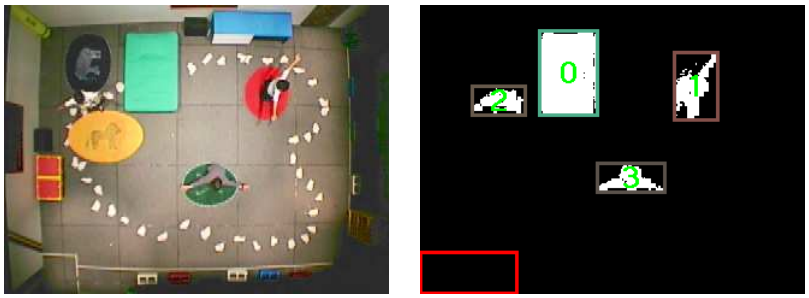

Figure 10: The left image shows a view with three people in the room from the overhead camera used for tracking.The right image shows the output of the tracking system, which is described and evaluated elsewhere[15]. All three people and the bed are being tracked as they move about. The box in the lower left denotes the room's door region, where all objects must enter and exit. The KidsRoom tracks up to four people and the bed, and people can enter and exit the space at any time.

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. In the leftmost image, the people on the bed are rowing on the left side of the bed. The center image shows rowing on both sides, and the rightmost image shows rowing on the right side. 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.

The rowing detection system demonstrates that it is possible to use narrative to constrain activity in an environment so that simple visual recognition techniques can be used to provide powerful feedback about activity in the space. However, once again, more constraints are placed on the story designer and the system must carefully check for situations where the context assumption has been violated. For example, the KidsRoom does check if someone leaves the bed during the rowing scene (responding with "Passenger overboard" messages and encouraging the person to get back on), but it does not check if the entire bed has been moved. If the bed moves, the algorithm will not perform well, because the system assumes that once everyone is on the bed rowing, the bed will not move.

7.3 Action recognition in the monster world

More sophisticated motion analysis is used during the dance segment of the monster world. There are four actions the system must recognize: crouching, throwing arms-up to make a "Y", 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; finally, 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. If one move matches well and matches better than the other moves, the system will signal the move has been detected. 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 characteristic of the background-difference blob. The "standing" blob shape for a person is initialized as soon as the person gets on the rug in the monster world. Once again the story is used to simplify the vision, because the assumption is made that when children first run to their rugs they will be standing, not crouching. Like the motion energy detection used for rowing, this simple technique is dependent upon being initialized in a specific contextual situation setup by the story.

Next, 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 changes 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

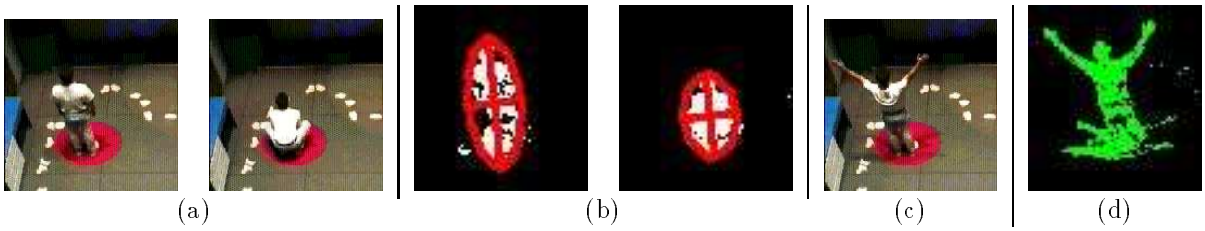

Figure 12: (a) A person performing a crouch move. (b) A person's background difference blob. Overlayed on top is an ellipse model of the blob. The first image shows a person in the standing position. The second shows the same person crouching. The difference in elongation of the ellipse model is used 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.

and crouching positions.

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[14] 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. A full description of the approach is available elsewhere[10]. In this method, 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 collapse the action over some interval of time down to a single image. The temporal extent (or range) 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.

7.3.4 Action recognition difficulties

All of the action recognition strategies make assumptions about the viewing situation. First, images of the children 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 powerful representation but requires that the lighting setup doesn't change between the training phase and run-time. Third, currently, the algorithms used in the KidsRoom are limited to recognizing actions containing single people, but there are many important actions in a multi-person collaborative space where there is physical contact. Examples include people dancing together and shaking hands. Actions that have subtle, small movement and that can be performed in a variety of ways (e.g. waving) are also difficult to detect using computer vision when people can move around a space as large as a the KidsRoom. In addition, the smaller the movement, the less likely it is that a camera will get an informative view.

7.4 Event Detection

In addition to recognizing large body motions of individuals, to provide interesting, reactive feedback most immersive environments will need to be able to detect other "events." 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?" "Have the kids just screamed?" and "Is someone near a particular object?"

The KidsRoom uses simple methods to compute answers to these questions. For example, the "in-a-group" detector gets the position of each person from the vision tracker and checks to make sure that every person is within some pre-determined distance of another person. The "on-the-bed" detector fires true when the tracker blobs for each person have merged into a single large blob which includes the blob for the bed (the bed can be moved and therefore must be tracked). The "just-screamed" detector checks if an amplitude threshold has been achieved over a short temporal window.

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 kid refuses to cooperate with the remaining children. In this case, a more robust detector might ignore the renegade given that he hasn't been following the rules for a while (e.g.using a "bad-kid" detector). 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.

Even detectors like "Is everyone moving around the path" are context-sensitive. For instance, it is not uncommon for people to pretend they are "creeping" as they move around the path. To the tracker, however, a group of people "creeping" together is barely mov-

Figure 13: Two of the dance move actions are recognized using a motion template matching method developed by Davis and Bobick[10]. The 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 right 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.

ing at all, and the "Is everyone moving around the path' detector might not fire. The system needs some method of adaptive temporal scaling.

Detection of context-sensitive events is a challenging area of future research in research on understanding action in interactive spaces. Handling unusual but possible cases will require special story, action, and control representations.

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 remote procedure call (RPC) connections.

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 changes how the room responds so that participants are guided through the narrative.

Vision processes, such as the object tracker, are continuously running and generating data. There are many situations that require an immediate response from the control program. 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 be frozen at its current point instead of continuing on as if there were still participants. Consequently, the control program can never stop and wait for some event at the expense of checking other perceptual input. The main control program is an event loop that continuously monitors the state of the room, checking all inputs as often as possible.

The control program keeps track of events using timers. Each event has a timer associated with it. When the event is activated, the timer is reset. The event timer can then be queried each pass through the event loop to see if the event has timed out. The most general event timer is simply used to time story events. For example, a timer is initiated for each short segment of the story. If the timer runs out, the narrative control program may then take some action like playing a narration or moving on to another part of the story. Timers are also used to control sound effects and narrations so that non-complimentary sounds don't play simultaneously. For instance, one background sound effect is a looping jungle noise. Each time it is played, a timer is initialized to the time of the sound clip. Each loop through the control program where there is a command to play the jungle sound, the timer is queried. If time is up, the sound is played again and the timer is reset. If the time is not expired, the sound is still playing and the control loop does not send the play command again. Since the control program is looping quickly, as soon as the sound stops playing it will play again, creating a continuous loop of forest ambient sound. Narrations are handled in a similar way, to prevent narrations from playing on top of one another or from playing too quickly one after another.

Event-based loops have been used as a control paradigm in the game industry and in commercial software like *Director* or *MAX*. However, dealing with real-time, physical interaction requires control structures more complex than those required in the typical keyboard-mouse situation, since 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. An alternative mechanism is to use the AI concept of planning to model the change in the states of the world [3]. Recently, Pinhanez, Mase, and Bobick [23] 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.

Tuning the narrative control for the KidsRoom is tricky due to unexpected interactions between various timers and suggests that more work in interactive control for immersive spaces is warranted. The timing problems we encountered will be discussed in section 9.3.6.

8.2 Music and sound control

The KidsRoom has an original score especially written for an interactive installation. The music consists of 50 short MIDI segments, many of which can be concatenated together 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.[7] Since the length of time any segment of the story might take depends upon the interaction of the people in the room, adjusting the timing of music segments that must end naturally at the end of a musical phrase has proven to be difficult. In these situations, the KidsRoom music tends to end at approximately the right time, but ideally the control architecture should be augmented with musical structures that allow for precise and rapid resolution of musical phrases.

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.

This model, while simple, has proven problematic. Our software platform has no mechanism for abruptly ending a sound that is currently playing. Once the sound starts streaming, it must play to completion. However, sometimes background sounds and narration need to be stopped at unpredictable times. For example, for the full effect, all sound should cease a split second prior to the mother's voice being heard in the monster world. This requires cutting music, narration, and background sound effects simultaneously. Unfortunately, such a simple sound playing mechanism does not give us this capability, and therefore some sounds play slightly longer than desired.[8]

[7] Computer control of the music is accomplished as follows. The control program sends a command to a process running on an SGI Indy workstation. That process translates the control command into a MIDI command, which is sent to a Macintosh computer running Studio Pro MIDI software. Studio Pro has been configured so that it accepts the MIDI commands and queues up particular segments of music. The SGI MIDI process listens to the Macintosh MIDI output commands, which are used to properly cue and interrupt music segments.

[8] While it is clear how to construct more responsive sound effects and music controllers, doing so requires modifying sound board drivers which is certainly possible but was too time consuming given our constraints when constructing the KidsRoom.

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 to give the KidsRoom a magical ambiance. Vision algorithms are disabled during the light transition segments of the story. The lighting is fully computer controlled, using a MIDI-controllable light board.

The most significant problem created by the lighting system is the vision systems inability to deal with small lighting changes. A single light blowing out during a run of the KidsRoom, partially darkening a corner of the room, can render the vision algorithms ineffective. Had we had more time for development of the KidsRoom, we might have altered the vision algorithms so that each world could have a slightly different color theme. This would require, however, that each world also have individual sets of vision thresholds for all of the algorithms. While this change would better fit the storyline of the KidsRoom, making the space more unusual and forest-like, the more complex lighting requirements would only increase the chance of minor lighting problems that disrupt the vision algorithms. Algorithms that re-acquire background models adaptively during the story might be required in immersive spaces with complex lighting. More useful would be a segmentation algorithm that is invariant to lighting changes altogether[17].

8.4 Animation Control

The KidsRoom uses layered, still-frame, cartoon-like animation sequences. These sequences are displayed by two image server programs, one for each screen. 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.

As animators know, people are sensitive to small inaccuracies in timing when watching human-like characters. In the KidsRoom, we do not need to synchronize partial monster movements with the movements of the people. However, if the system used more complex monster actions, and especially if these actions needed to be synchronized precisely with the action of children in the space, added complexity would be required of the control program.

8.5 Process Control

The KidsRoom control program receives information about the state of the room and controls the output of the room using ten server programs. The sensor servers are the object tracker server, the motion detector server, the action recognition servers (2), and the scream detector server. The output servers are the directional sound server, the music server, the lights server, and the display servers (2).

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[7], 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.

9 Design of the interactive narrative

This section is divided into three parts. The first part continues to elaborate on our design decisions. Next, we further describe constraints we encountered when building an interactive room using unencumbering sensing technology. Finally, we discuss some general observations and failures.

9.1 Design decisions

In section 4 we outlined the primary design decisions we made before constructing the KidsRoom. In this section, we discuss those goals and others.

9.1.1 Action in the physical space

Our primary goal, in the spirit of the work reviewed in section 2, was to construct an immersive environment where action and attention was focused primarily in the room, not on the screens. Further, we wanted the room to be a physical space, where multiple children could play much as they do in real life, only within a rich environment that would watch what they do and respond in natural ways.

We believe the KidsRoom achieves these goals. Children are typically active when they are in the space, running from place to place, dancing, and acting out rowing and exploring fantasies. The kids are interacting with each other as much as they are interacting 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. When action on the screens captures the attention of the children, the children are still focused partly on the real space as they perform large body actions and observe and communicate with other children.

9.1.2 Invisible sensing

One way to make it easier for the focus of activity to remain in the physical space is to free a room's occupants from wearing sensors, head-mounted displays, earphones, microphones, or specially colored clothing. We were committed to using computer vision tracking and action recognition software to allow any person to easily enter and exit the room at any time naturally; in the KidsRoom there are no requirements of people who enter the space except that they must enter one at a time. Cameras and a single, stationary microphone are the only forms of input – and all are completely non-intrusive. Vision can differentiate actions that floor sensors cannot, such as flapping and making a Y pose, and we believe that in the long run it will be less expensive and more practical to outfit a space with cameras and possibly a few other types of sensors rather than without cameras and with hundreds of object sensors.

Further, occupants in the room (particularly young children) are not even aware 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 increases 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.

9.1.3 Perception and Story

The immersive power of a compelling storyline cannot be undervalued when constructing a room like the KidsRoom that integrates technology and narrative. The presence 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-designed 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.

An immersive space is most compelling 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. The room is and appears to be responsive. Monsters follow the dance moves of the children. The room transforms the instant that everyone jumps on the bed. Monster growling stops when children hide behind the bed. Monsters react when children scream loudly, and the narrator reacts when they don't. The kids can move and steer their "magic boat" using natural body gestures. The room responds when they are near particular pieces of furniture. Upon entering the room, the children quickly realize that what they do makes a difference in how the room responds. This perceptual sensing enhances and energizes the narrative.

Conversely, 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. In the KidsRoom, nearly all of the algorithms are dependent upon the story to provide constraint. The boat rowing motion energy algorithm works because the story ensures the people are going to stay on the boat during the scene. The monster dance scene uses the story to ensure that each camera has a non-occluded view of a child performing actions and that no other kids will move over and interfere.

A goal that was critical to obtaining the immersive feel of the kidsroom was to *naturally* embed the perceptual constraints into the storyline. 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." A less obtrusive choice, however, is to embed the constraint within the behavior naturally assumed of the characters. The *monsters* tell the kids to stay on the rugs and that there can be only one kid 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 ques-

tion why they need to engage in some particular behavior.

9.1.4 Multiple, collaborating people

Another design goal was to create a space that could handle the interaction of multiple people. Spaces with more than one person are a challenge for most computer vision algorithms and interactive systems, but people are clearly drawn to experiences where there is group contact and interaction. 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. The focus of attention in a room is more likely to remain in the physical space if there are multiple people in the environment. If they are 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. This is clearly the case in the KidsRoom. Much of the fun is being in the room with other people and "playing" together. 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.

Keeping action focussed in a real space with multiple people will encourage collaborative activity, and the KidsRoom was designed with this in mind. To participants, the room appears interested in group activity: "Is everyone on the bed?" "Is everyone in a group?" "Is everyone rowing on the correct side?" "Is everyone on the path?" In practice, the group activity recognition techniques are simple but should be more context sensitive, because the more people that participate in an immersive experience, the more likely there will be someone who engages in some unpredictable behavior. Similarly, the responses the system must have available grows in number.

In addition to narrative complexity, adding more individuals to a space increases the complexity of visual recognition. The major issue is occlusion and camera line of site. As more multiple enter the space, some people will block the view of others from particular cameras. In the monster world of the KidsRoom, a large space was required to get a non-occluded view of just two children. The KidsRoom does most of its checks for group activity using positional information. However, there are currently no mechanisms that allow recognition of large-body group activity such as two children dancing holding hands and two people shaking hands.

The KidsRoom's use of a multi-person tracker and special constraints built into the story to handle recognition of group activity made it possible to build an environment that can handle up to four people.

9.1.5 Real objects

We strove to develop a system where people manipulated real objects in a real space. Although the KidsRoom is populated with real objects, with the exception of the bed, they are all immobile and can't be manipulated. The bed can be rolled around, jumped over, and hid behind, and it is a critical part of the narrative.

The two major obstacles to incorporating more objects are (1) tracking them accurately as the space becomes more visually cluttered, and (2) expanding the control mechanisms and narrative branching so that the exponential number of problematic scenarios created by the objects can be handled.

The KidsRoom tracking algorithm sets a limit of four people and one bed in the space because when more participants or large objects enter the space becomes visually cluttered. Consequently, the people and objects tend to merge into larger and larger blobs in the imagery, and the blobs are computationally difficult to divide into components. Small objects, conversely, won't cause additional object merging but are difficult to track because they are easily occluded and often have small visual cues that are difficult to track. More research on tracking multiple people and objects in small enclosed spaces is warranted in order to develop algorithms that are useful in typical household and office environments where there can be hundreds of manipulatable objects in just a few feet of space.

The second and 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 since the ways in which objects may be used (or misused) increases. Further, the ways in which objects may interfere with perceptual routines will grow in number. 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 immersive the entire system feels. Graceful degradation, discussed next, will become increasingly important as more objects are added to a space.

9.1.6 Graceful degradation

No matter how well an interactive story like the KidsRoom 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 completely destroy the entire immersive experience. When perceptual algorithms fail the behavior of the entire room will degrade 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 doesn't appear that the room is "confused." However, if the monster was saying "Great crouch" and the kid knows she is spinning, the room

appears less responsive.

Similarly, we tried to minimize the number of story segments that required a specific, singular action on the part of all the participants. For instance, to get a particular piece of furniture to talk, only one child needs to be near it. If all children were required to do it they might never discover how to make the furniture talk.

When a particular action is required, the story components are designed so that the system can handle a problematic participant like a grouchy child or a mischievous graduate student. In such cases, the system tries to acknowledge that someone is doing something that they shouldn't be, but if the person continues to cause trouble, the system will continue progressing through the story anyway.

9.1.7 Visual and audio feedback

Once the KidsRoom detects an action occurring in the space, it must alter the environment of the room in some way both to acknowledge the event detection and to drive the narrative forward. 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 storybook look and video would consist of simple, still-frame animations of those images. Although the animations are simple, most consisting of fewer than five frames, small effects can make the characters more interesting. For example, when the monsters are standing still, they blink, giving them a more life-like appearance. During much of the KidsRoom experience, the video screens are used as a mood-setting backdrops and are not always the center of the participants attention.

Audio is the main form of feedback in the room, which has several advantages over video. Realistic or semi-realistic three-dimensional video animation, even of cartoonish monsters, is difficult to do well. It is much easier to obtain rich and realistic sound effects and narration. Perhaps most importantly, 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. However, during the river world and part of the monster world, where video feedback is more important to the storyline, the children are forced to stare at the screens. Finally, 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.

One of our goals was to create an "imagination space" – a place that stimulates creative play by creating a rich environment but one where much detail is left to the child's imagination. Audio feedback is a powerful tool because it can be used to inform participants of story events while leaving gaps that children can fill in as they like, much as they do in their everyday play.

9.1.8 Story consistency and immersion

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 8. Some of the furniture characters in the first world have the same voices as the monsters in monster world. The artwork has the same storybook motif in all four scenes. Some objects in the room on the shelves become part of the forest world backdrop during the transformation.

Although we are uncertain if kids and adults consciously notice these minor story points, attention to such detail minimizes the chance that some awkward story plot will jar some participant out of the immersion that the space is continually trying to generate. Careful attention was played to larger narrative details as well. The primary narrator, the room, always speaks in child-like rhymes.

Much like film makers, we strove to achieve a "suspension of disbelief" in the audience members. Thus, as in any well-written story, interactive or otherwise, attention to detail is required. What we envisioned when writing the story and developing the space was to create an engaging, interactive, theater-like experience. We provided a stage set – familiar, but larger-than and more colorful than real-life. We provided a story in which participants became actors. We provided theatrical lighting and musical accompaniment. We provided a narrator guide whose rhyming couplets were simultaneously dramatic and encouraging. We provided a set of characters – talking furniture and dancing monsters – primed for adventure and a certain amount of hilarity. All of these elements combine to yield an engaging and theatrical, but participatory, experience in which a group of children are able to immerse themselves in the momentum of the story and to contribute to that momentum as well.

9.2 Constraints

A major challenge when designing the KidsRoom was trying to minimize the impact of our technical constraints on the development of an interesting story. This section expands on a few constraints that have not been discussed previously.

9.2.1 Lighting

Lighting is often used effectively in theater to set a mood and highlight an important event. In fact, lighting has been used extensively in much of the previous work on immersive environments discussed in section 2. However, systems that use visual input have been constrained to operating in relatively bright, constantly lit environments.

Unfortunately, bright white lighting, ideal for computer vision, requires a serious compromise on the part of the designer of an immersive space. Controlled lighting can be an especially effective way to influence the physical world in a natural, but powerful way. In the KidsRoom, the only major lighting effect, the blackout and colored lighting during the transformation from bedroom to forest worlds, clearly captures the attention of both children and adults. Prior to the blackout, after running around the room "talking" to furniture, children will often be rather hyper. Only moments later, however, after passing a few seconds in the dark on the bed, they are typically more calm.

Large projected displays appear dim when placed in bright spaces. Although we obtained satisfactory brightness using light blinders, the forest, river, and monster world would seem more mysterious if the room were darker and the forest screens brighter. Unfortunately, bright screens reflect more light back into the room which sometimes degrade performance of color-based vision algorithms.

9.2.2 Screens and vision

One serious problem we encountered when developing the space was the problematic interaction between video displays and computer vision algorithms. Many real-time computer vision algorithms require the use of background subtraction, as do all of the algorithms used by the KidsRoom. However, background subtraction assumes that the background is static. Large video screens with changing images violate this assumption. 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 with two walls that are screens. As shown in Figure 1, the two wall-mounted cameras are placed high above the room so that when an adult stands on the red or green rug, the person's entire image is over the floor, not the changing screen (see Figure 3).

One problem for those developing immersive environments is that large screens are an effective way to provide feedback. Ideally they should be distributed around the room. However, in an environment like the cave[8], where all four walls are screens, few camera angles would be suitable for our vision algorithms and those that were would not show useful profile views of people near the screens. This limitation must be overcome before vision algorithms can be used to their full potential in such immersive spaces. One solution being studied is using real-time stereo vision for background segmentation[17].

9.2.3 Physical layout

Significant thought and testing was required to physically layout the KidsRoom space, and perception-based requirements had to be carefully considered when developing the KidsRoom narrative. The problem was finding a way to mount the video cameras so that the action that needed to be recognized could be interpreted by the vision systems. Even in a space as large as 24 feet by 18 feet with a high ceiling camera placement was tricky. The top-down camera makes it possible for the tracking algorithm and motion-energy algorithm to operate without explicit reasoning about occlusion. However, the tradeoff is that our algorithms were developed with the assumption that a camera could be placed over 20 feet above the room, which is out of the question for most normal indoor, household spaces. The top-down camera imposes the weakest constraint: we cannot easily hang an object, lighting fixtures for example, above the space since it will block the camera's view.

The other two cameras used for recognition on the red and green rugs severely constrain the layout of the room. As discussed in section 9.2.2, images of people where video is directly behind the person in the camera view cannot be used due to the use of background subtraction algorithms. This constrains the rugs to be a certain distance from the screens since the cameras get a poor view of each person's silhouette if they are placed too high on the wall. In addition, the rugs had to be placed far enough apart so that when two people spun around with arms out, the silhouettes would never touch in either image.[9] Finally, the rugs need to be near the screen where the monsters are "standing." These constraints dictated precisely where the front two rugs *had* to be placed. The bed and the back two rugs were also under tight positional constraints.[10]

Every space-related decision required careful consideration of imaging requirements. All objects other than the bed had to be carefully fastened. Rugs and carpet had to be short-haired, not the least bit shaggy, to prevent the background from changing as people moved around. Objects were painted in flat paint to minimize specularity from the bright lighting and transfer of color from objects to people by reflection. Any object in the line-of-sight of any camera could not have specular metal parts. Even the sheets on the bed were tightly fastened to the frame.

The use of narrative in the KidsRoom made it possible for us to overcome these limitations by carefully integrating visual constraints into the story constraints. However, these type of stringent requirements are clearly limiting for the interactive story designer.

9.2.4 Unencumbering audio

Constraints on audio processing currently restrict the type of stories that could be developed for the KidsRoom. The first constraint is lack of effective sound localization. The four speakers in the KidsRoom pro-

[9] The action recognition technique used by the KidsRoom can use multiple viewpoints of the same action to make recognition more robust[10], although in the actual KidsRoom implementation only one view was used for each move. Using multiple views further constrains the room layout.

[10] If the back rugs were too close to the screens, the top of an adult could end up overlapping with the image for a person on a front rug. Similarly, the bed needed to be out of the line-of-sight. The back rugs needed to be far enough apart so that the tracker could be certain that one person was on each rug. If the rugs were too close, the image blob of the two people would merge leaving the tracker uncertain of each person's exact position.

vide 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, because the effect is destroyed if the sound is not perceived to come from the object. One solution to this problem is to populate the room with many small speakers concealed carefully in walls and some large devices so that sound can really can come directly from a "talking" object.

The second audio issue left for future systems is the use of voice recognition. Unlike some other immersive environments that require portable microphones[7], the KidsRoom has no mechanisms for understanding speech. Although large-vocabulary speaker-dependent and small-vocabulary speaker-independent continuous speech recognition systems are available, they generally require the user to wear a noise-canceling headset microphone. Adding unencumbering speech recognition capability to a space like the KidsRoom with loud music, loud sound effects, and loud children remains a significant challenge. Clearly, the ability to recognize some speech would expand the types of interactive rooms that can be constructed, since some feedback from the participants to the room is accomplished most naturally using speech not visual gesturing.

9.3 Observations and failures

There were some important issues that we failed to consider in the design phase of the KidsRoom that are important for developing other immersive spaces, particularly those for children.

9.3.1 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 capability; the KidsRoom guesses how long each activity should take. There are ad hoc timing adjustments that scale based upon the current number of people in the space.

In general, the more children that are in the space, the more fast paced the room appears to be, since as soon a single 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. The child often requires more intervention from the system to guide him or her through the experience.

Occasionally, a single problematic kid can lead all the other children to engage in behavior that the room did not predict. The room might handle these situations better by individually recognizing and encouraging behavior that is most appropriate for the developing story to try and pull the deviating kids back from whatever else they are doing.

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

9.3.2 Exploratory vs. linear spaces

In our initial design of the KidsRoom, we planned to create a mostly exploratory space, modeled somewhat on popular non-linear computer games like Myst[21]. 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, as if the mirror broke. Moving near a real or virtual window and putting an arm towards the window generated a loud crash of glass. Stepping on the rugs with animal pictures elicited the corresponding animal noises: a ribbit for the frog, a roar from the lion.

Our hope was that children would enter, one at a time, 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 that way. When we brought in some children for testing we found that they did not understand that they were causing the sounds – there was too much going on. Even when only *one* child was in the space, the cause and effect relationship was not explicit enough for the children to grasp. The same was true for some adults.

Part of the problem was that there is a slight lag in the perceptual system. After stepping on a rug, a half second might pass before the sound is heard. The larger problem, however, was spatial accuracy. We could only tell when a child was close to the window, so the system could not distinguish between a kid standing next to the window making smashing gestures and a kid that just ran right by the window. Even if our sensing had been more precise, however, we believe that with multiple children there would be too much happening in the environment to make the space engaging. We completely revised the first world, using the more linear talking furniture model.

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, and we modified the forest world so that there was more narrative structure by constructing the path and, essentially, having the characters tell the kids what they should do.

Clearly, it is easier for an immersive environment to respond to action taking place in a linear story because the number of action possibilities at any moment is more constrained. Moreover, other authors have observed that exploratory, puzzle-solving

spaces can sometimes make it difficult for adults to immerse themselves in an interactive world[12, 9]. 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.

9.3.3 Iterative development and testing with children

We underestimated the need for rapid, iterative development of each component of the system and concept testing with children. We were limited to a just a few sessions, each of which had to be short since the kids would tire quickly of any sort of repetitive testing of a particular component of the room. More problematic was that once the kids went through the room once, we were never certain how much further testing was influenced by the initial runs. Our first test with children made it clear that children needed to be a part of the testing process. We seriously revised our exploratory design. Later tests proved to be just as valuable.

One problem we encountered was that small changes in the interactive nature of the room often meant adjusting several subsystems: the vision programs, the control program, the storyline, the narrative, and sound-effects. Everyone developing the system was not physically located in the same space and not available at the same time. Therefore, writing, recording, editing, and integrating new narrative and sound became a bottleneck preventing rapid tuning of the system. Some of the voice talent was located off-site, which slowed the process further, and even though temporary narration was used in development, the interaction during the limited time we had with children was not as effective as it might have been.

In future endeavors, we would make a more serious effort to follow a studio model where every effort is made to ensure that story modifications can be made quickly and rapidly and integrated into the full system. Further, more thought should be put into how to organize and maintain the hundreds of sound and image files that were required for a space of the KidsRoom's complexity. Such a setup would make testing, and especially testing with children, more informative.

9.3.4 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.

First, the story must take into account the children's behavioral "momentum." The KidsRoom is capable of making children somewhat hyperactive. By the end of the first bedroom world when the furniture all start loudly chanting the magic word, the kids are often running energetically around the room. Next the kids end up on the bed, scream the magic word loudly, and the transformation occurs. Then the kids are in the forest world, and they are expected to explore. However, the transformation typically calms them down and their tendency is often to stay right where they are on the bed. We found through testing that it takes a fairly direct instruction (e.g. "Follow the path..."), sometimes repeated several times, to get them to start moving again. We failed to anticipate this behavior and had to later modify the storyline to improve the interaction. 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. Again, we had to modify the narrative so that it repeats some critical instructions more than once. Ideally, we should have built attention-grabbing narrative into the storyline for every critical narration. Since most children have not been in any type of "active room" before, preparing them briefly before they enter by telling them to be attentive to the room's instructions has proven effective.

Finally, children need to clearly understand the current task. The less certain they are of what to do, the more unpredictable their behavior becomes. The kids seem more inclined to wait for things to happen than to explore and try to make things happen, so when the room didn't provide a goal, they would tend to sit on the bed and talk or just look around. The kids seemed to enjoy the experience the most once the system was modified so that there was always a clear task and when those tasks changed quickly.

9.3.5 Avoiding repetitiveness

One way to break the immersive experience is to have the system start repeating the identical narration over and over as it tries to encourage some behavior. Unfortunately, in a space like the KidsRoom especially built to encourage children to physically move around, instructions do need be repeated because sometimes they aren't heard or are ignored. For example, in the dance segment of the monster world, the control program continually checks if someone has gotten off a rug or if two people are on the same rug. If someone gets off a rug more than once, narration is needed, but playing the same narration as just played moments before makes the system sound broken and feel less immersive.

One solution we developed was to use two different narrators. The main narrator has a deep, male voice. 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. Hints are used throughout the KidsRoom and were effective at – in fact, required for – successfully improving our narrative where our original storyline was problematic. For example, when the kids don't put the bed back in its proper position after the river world, they will hear several hints about putting it back and a hint telling them where to put it.

Hints do not solve the repetition problems, however. Most KidsRoom hints have several variations so that if they must be repeated they sound slightly different each time. Most narration, though, due to time constraints, does not. It became clear that practically every narrative instruction and certainly every hint needs *at least* three different variations so the control program can repeat instructions without breaking the immersive feel. Ensuring that the system avoids repetitive narration adds some additional complexity to the control code.

9.3.6 Sensitivity to timing

As discussed earlier, people are sensitive to small timing problems and the control architecture proved to be sensitive to small changes in timing values. As a result, multiple people in a space increases the number of possible situations and responses that are required, thereby making the problem more difficult.

A typical example is as follows. Several kids are on the bed rowing down the river. A narration begins playing, that will take 4 seconds. During the first second, someone gets off the bed. However, the man-overboard narration (which takes 4 seconds as well) can't play because of the first narration. The system doesn't respond. If the person gets back on the bed quickly, he or she may not realize the room recognized what happened. Just as problematic is if the person stays off the bed for several seconds. Then the overboard narration plays, only 3 seconds too late. The system feels sluggish, even though the recognition systems could detect the overboard event almost instantaneously – the room just had no way to respond. This problem is encountered in many situations, particularly those where several different tests are active simultaneously.

In general, our narrations are far too long, often taking several seconds to read. Since we have no way to interrupt narration, it is difficult to adjust the control timing to avoid sluggish response. Shorter narrations and short sound effects would help, but not solve the problem. Even if we could cut the sound of a narration mid-phrase, such unnatural cutoffs are likely to break the suspension of disbelief.

Although the timer controls are fairly straightforward to program, it is time-consuming to tune the narrative control system so that the room responds naturally throughout each stage of the story. For example, in a given segment of the story, it is possible to program timers so that the interaction feels "right" when there is one person in the room following the room's instructions. However, when three people are in the room and they are not cooperating, additional narrations may be required since room participants may perform actions differently and take longer or shorter amounts of time. Consequently, the timing can be thrown off and the room can start to feel unresponsive. Further, the more people that are in the room, the more situations the control program must handle. Before long, the number of different situations has forced the programmer to add many special timing conditions that increase the complexity of the program and make it time-consuming and tricky to adjust. Handling such problems while maintaining a feel of quick responsiveness without generating repetitive narration has required a large amount of effort and experimentation. Some of the authors are working on methods that represent temporal events more naturally[25] that might potentially help identify tricky timing situations automatically or semi-automatically.

The lack of a systematic method of checking for timing inconsistencies 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 over and over. 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.

9.3.7 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. 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 kids like to scream, 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 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. The downside, however, is that the amount of narration required is at least three times what it was before.

Similarly, there are times when the system is dealing with uncooperative participants and, despite several attempts, has not elicited the desired activity. This occurs, for example, if someone refuses to get on the bed in the bedroom or monster worlds even after the mother is heard telling them to do so three times (with three different narrations). In this case, we decided to have the narration move on anyway, ignoring the trouble-maker. However, it is important to explicitly *acknowledge* that the system understands something is wrong but is ignoring the problem. One example in the KidsRoom is when nobody shouts "Be quiet" to the monster. The narrator says, "Well, I can't say *that* was a very loud shout, but perhaps the monsters will figure it out." However, currently the KidsRoom does not have a similar narration if someone refuses to get on the bed, hence, the story moves on without explicitly saying that it knows someone is

still off the bed and an opportunity to demonstrate responsiveness is wasted.

9.3.8 Perceptual expectation

Given the technological limitations on unencumbered sensing in immersive environments like the KidsRoom, the narrative must be carefully designed so that it does not setup any perceptual expectations 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 to figure out what the room understands and what it doesn't, is more likely to feel more broken and unresponsive than immersive.

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 can't recognize (i.e. the dance moves in the monster land).

10 Summary

The KidsRoom went from whiteboard sketches to a full 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 more complex spaces in the future.

The KidsRoom is an immersive environment where the action stays focussed in the real physical space that is augmented using images, music, narration, and sound. The system is fully-automated and has been tested with children from ages six to sixty. Up to four people, unencumbered by sensing devices, can enter the space and, without instruction, embark on a fantasy adventure. Computer vision action recognition technology is used to sense and interpret the environment of the room, and control mechanisms have been built that respond to the activity of room participants. Perceptual recognition makes it possible for the room to respond to the children and for the children to be aware of this responsive behavior. Further, the narrative context of the story makes it easier to develop context-dependent (and therefore more robust) action recognition algorithms. The room recognizes and responds to individual actions like "spinning" as well as group collaborative activities such as "rowing" and "walking in a group."

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 coordinator. 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 and handled most narrative recording. 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), 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 *Proceedings 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. In *Advances in Human Computer Interaction*, volume 5, chapter 2. Ablex, 1995.

[5] A.F. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. In *Royal Society Workshop on Knowledge-based Vision in Man and Machine*, London, England, February 1997. To appear: Proceedings of the Royal Society.

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

[7] M.H. Cohen. Building brains for rooms: designing distributed software agents. Technical report, MIT AI Lab, Cambridge, Massachusetts, 1997. http://www.ai.mit.edu/projects/hal.

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

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

[9] 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.

[10] J.D. Davis and A.F. Bobick. The representation and recognition of action using temporal templates. In *Proc. Comp. Vis. and Pattern Rec.*, San Juan, Puerto Rico, June 1997.

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

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

[13] J. Glos and M. Umaschi. Once upon an object...: computationally-augmented toys for storytelling. In *Submitted to Int'l Conf. on Computational Intelligence and Multimedia Applications (ICCIMA)*, 1997.

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

[15] S.S. Intille, J.D. Davis, and A.F. Bobick. Real-time closed-world tracking. In *Proc. Comp. Vis. and Pattern Rec.*, pages 697–703, San Juan, Puerto Rico, June 1997. IEEE Computer Society.

[16] H. Ishii and B. Ullmer. Tangible bits: Towards seamless interfaces between people, bits, and atoms. In *Human Factors in Computing Systems (CHI)*, pages 234–241, Atlanta, Georga USA, 1997.

[17] Y. Ivanov, A.F. Bobick, and J. Lui. Fast lighting independent background subtraction. Technical report, M.I.T. Media Laboratory Perceptual Computing Section, August 1997.

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

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

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

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

[22] A. Pentland. Smart rooms. *Scientific American*, April 1996.

[23] Claudio S. Pinhanez, Kenji Mase, and Aaron F. Bobick. Interval scripts: A design paradigm for story-based interactive systems. In *CHI'97*, pages 287–294, Atlanta, Georgia, March 1997.

[24] C.S. Pinhanez. Computer theater. Technical Report 378, M.I.T. Media Laboratory Perceptual Computing Section, May 1996.

[25] C.S. Pinhanez and A.F. Bobick. Human action detection using PNF propogation of temporal constraints. Technical Report 423, M.I.T. Media Laboratory Perceptual Computing Section, May 1997.

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

[27] Walt Disney Productions. *Bedknobs and Broomsticks*. 1971. Movie.

[28] M. Resnick, F. Martin, R. Sargent, and S. Silverman. Programmable bricks: toys to think with. In *IBM Systems Journal*, volume 35(3-4), pages 443–452, 1996.

[29] M. Sendak. *Where the Wild Things Are*. Harpercollins Juvenile Books, 1988.

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

[31] M. Torrance. Advances in human-computer interaction: the intelligent room. In *Working notes of the CHI 95 research symposium*, Denver, Colorado, 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*, Montreal, Canada, August 1995.

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

[34] R. Want, A. Hopper, V. Falcao, and J. Gibbons. The active badge location system. In *ACM Transactions of Information Systems*, volume 10, pages 91–102, 1992.

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

[36] P. Wellner. Interacting with paper on the digitaldesk. In *Communications of the ACM*, volume 36, pages 86–96, 1993.

[37] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. In *Proc. of the SPIE Conf. on Integration Issues in Large Commercial Media Delivery Sys.*, Oct. 1995.