Design Decisions for Interactive Environments: Evaluating the KidsRoom

Aaron Bobick, Stephen Intille, Jim Davis, Freedom Baird, Claudio Pinhanez
Lee Campbell, Yuri Ivanov, Arjan Schütte, Andrew Wilson

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

Abstract

We believe the KidsRoom is the first multi-person, fully-automated, interactive, narrative environment ever constructed using non-encumbering sensors. The perceptual system that drives the KidsRoom is outlined elsewhere(Bobick *et al.* 1996). This paper describes our design goals, successes, and failures including several general observations that may be of interest to other designers of perceptually-based interactive environments. [1]

Introduction and Goals

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.

In this paper we describe observations made as we constructed and tested one such space: the KidsRoom. Our hope is that the lessons we learned will be valuable to others designing such perceptually-controlled, interactive and immersive spaces.

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

- To keep action in the physical space.
- To use vision-based remote sensing to encourage unencumbered, natural interaction.
- To construct a system that worked effectively with multiple people.
- To reduce the brittleness of perceptual sensing by allowing the system to both use and manipulate context.
- To create a space that was immersive, engaging, and fully-autonomous and that could be interacted with naturally, without prior instruction.

The idea of constructing a children's "imagination playspace" immediately addressed most of these concerns: multiple children being active in an engaging activity. The goal of exploiting and controlling context was accommodated by embedding the experience in a narrative environment, where there was a natural storyline to drive the situation.[2] Finally, by building a large scale room, we could insure an immersive experience and provide the necessary viewing conditions to allow the use of visual sensing.

The playspace, story, and interactivity

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. Some effects are particularly loud and can vibrate the floor. Finally, there are three video cameras and one microphone installed. The room contains several pieces of real furniture including a movable bed, which is used throughout the story. Figure 1 shows a view of the complete KidsRoom installation. The KidsRoom experience lasts 10-12 minutes, depending upon how the participants act in the room, and it was designed primarily for children ages 6-10 years old. Throughout the story, children

[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.

[2] An example of a non-narrative space would be a kitchen that watches what people do and responds by offering advice.

Figure 1: The KidsRoom is a 24 by 18 foot space constructed in our lab. This image shows the entrance to the space, the two projection screens, the computer cluster outside the space, and the position of two of the cameras used for sensing.

interact with objects in the room, with one another, and with virtual creatures projected onto the walls. The actions and interactions of the children drive the narrative action forward. And, most importantly, the children are aware that the room is responsive.

The KidsRoom was inspired by children's stories in which children are transported from their bedrooms to magical places. Here we briefly describe the experience (see (Bobick *et al.* 1996) for a detailed description of the worlds and the interaction of participants).

In the first world, children enter the bedroom being told only to "ask the furniture for the magic word." A tracking algorithm(Intille, Davis, & Bobick 1997) is used to determine when children are near pieces of furniture, each which has a different "personality." 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 "frog rug" may send send them to yet another piece of furniture. If the children get confused and don't go to the correct place, the system will respond by having some furniture character call the children over, "Hey, over here, it's me, the yellow shelf!" Eventually, the children discover the magic word. 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.

After the kids yell the magic word loudly,[3] the room darkens and the transformation occurs. As colored-lights flash and mysterious music plays, images on the walls fade to images of a cartoon fantasy forest land. Simultaneously, a deep-voiced narrator — the voice of the room — says, "Welcome to the KidsRoom. It's not what it seems. What you might see here are things dreamt in your dreams." The narrator speaks in rhyme throughout the story.

As the lights come up, the narrator tells the children that they are in the "forest deep," that monsters are near, and that they must follow the path to the river. A stone-path is marked on the floor of the room and the children quickly assume that is the path they should follow. The room provides encouraging narration if they don't do so, and instructs them to to stay in a group and to remain on that path. If they deviate from the path "hints" (discussed below) are given to induce the behavior. Later, the children must hide behind the bed from roaring monsters.

After a short walk, the children reach the river world. One view shows the river progressing forward and the other view shows the moving riverbank. In this world the children get on the bed and control it to avoid obstacles in the river by making rowing motions, recognized by a motion detection algorithm(Bobick *et al.* 1996) as shown in Figure 2-a. The context of the story is used to control the interaction. For instance, if someone gets off the boat, a splashing sound is heard and the narrator yells, "Passenger overboard" and encourages the child to get back on the bed.

Eventually, the children reach the monster world. In this world they land the "boat" by pushing the bed and then still-frame animated creatures teach the children four dance moves. The monsters, shown in Figure 2b, are larger than the children and have a friendly, goofy, cartoon look. When the appear, the kids must yell to quiet them. The kids learn the dance moves as they stand on rugs in the room. Computer vision action recognition algorithms(Davis & Bobick 1997; Bobick *et al.* 1996) are used to recognize when children do the moves, and the characters give feedback like "Yo! Kid on the red rug, you dance like a pro!" Eventually, if the children do one of the four moves they learned, the monsters will copy (see Figure 2-b). The story ends with the mother once again breaking in and telling the kids to go to bed, at which time the room transforms back to a bedroom (see Figure 2-c).

Achieving project goals

We review the goals presented in the first section of this paper, considering not only how well the goals were achieved but also the implication those goals had on the development of perceptual algorithms and the overall success of the project.

Real action, real objects

One of our primary goals was to construct an environment where action and attention was focused primarily in the *real physical environment*, not on the virtual

[3] There is no speech recognition capability, with only the volume of sound being measured. In no run of the KidsRoom did the children ever yell anything but the magic word.

(a) (b) (c)

Figure 2: (a) A child and mother row the boat together. (b) Children dancing with monsters. (c) Children "going to bed."

screen environment. We wanted the room to be a physical space, where 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 this goal. Children are typically active when they are in the space, running from place to place, dancing, and acting out rowing and exploring fantasies. 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.

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

The KidsRoom tracking algorithm(Intille, Davis, & Bobick 1997) sets a limit of four people and one bed in the space because when more participants or large objects are present the space becomes visually cluttered. Small objects, conversely, are difficult to track because they are easily occluded and often have small visual cues. Better techniques or even paradigms are needed to track multiple people and objects in small enclosed spaces such as typical household and office environments where there can be hundreds of manipulatable objects in just a few cubic feet of space.

The second and perhaps more serious obstacle is narrative control. As more objects are added to a space, the behavior of the people in the space will become less predictable since the number of ways in which objects may be used (or misused) increases. Further, objects become more likely to interfere with perceptual routines. 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 and sentient the entire system feels.

Remote visual sensing

In the KidsRoom there are no incumbrances or 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. Remote sensing facilitates the activity remaining in the physical space, not requiring users to wear sensors, head-mounted displays, earphones, microphones, or specially-colored clothing.

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 of and between the people in the space. Second, physically wiring every object in a complex space with several different devices makes modification to the space and objects a hardware procedure, increasing experiment cycle time and expense. 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. Vision sensing can provide both the global and local information required to interpret activities.

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.

As a practical implication, using visual sensing means that adding a new element to the story that requires a new action to be sensed is not a hardware operation. One installed, the sensing infrastructure remained stable eliminating potentially combinatorially

difficult networking issues.

Multiple, collaborating people

Previous work in computer vision and fully-automated interactive spaces has primarily considered environments containing only one and sometimes two people(Maes *et al.* 1994). However, immersive spaces for living environments or office environments will require systems that understand the interaction of multiple people. Further, the focus of attention in a room is more likely to remain in the physical space if there are multiple people in the environment since, if unencumbered by head-mounted displays, people will naturally communicate with each other about the experience as it takes place, and they will watch and mimic one another's behavior. This is clearly the case in the KidsRoom.

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. 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.

However, being responsive to the actions and interactions of multiple people required greater use of context specific techniques. The more people that participate in an activity, the less likely one can accurately describe the appearance of that action. However, cues that are indicative of behavior can be safely used within a context. An example in the KidsRoom is rowing, when it is possible to use context to enforce constraints needed to recognize a complex group activity (i.e. rowing) with a simple global motion detection algorithm. Without using the context of the story to encourage certain behavior (e.g. getting everyone on the bed) and allow the use of context-selected perceptual methods the multi-person activity would be much more difficult to recognize.

Multiple people also increases the complexity of visual sensing because of the presence of visual occlusions. As more people enter the space, some 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 all of its recognition of group activity using positional information from the overhead camera view in which occlusion does not occur. 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.

In addition to complicating visual tasks, the presence of multiple people increases the narrative complexity. 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?" However, the more people present the greater the likelihood one person will engage in an unpredictable or inappropriate manner. The narrative control needs to anticipate such situations and be able to respond non-repetitively when necessary.

Exploiting and controlling context

Computer vision techniques are still brittle. Given the difficulty of designing robust perceptual systems for recognizing action in complex environments, we strove to use narrative to provide context for the vision algorithms. Most of the vision algorithms are dependent upon the story to provide constraint, for instance the boat rowing mentioned in the previous section.

Another example is the monster dance scene where the story ensured that each camera has a non-occluded view of a child performing actions. Potentially interfering children are cajoled by the monsters to stand in locations (on colored rugs) that do not interfere with the sensing. The advantage of an active system over that of a monitoring situation is the opportunity to not only know the context but to *control* it as well (see (Bobick *et al.* 1996) for other examples).

Immersion, story, and consistency

We wanted an environment that was truly immersive, completely perceptually and cognitively engaging, and where participants did not need to ask for outside help. The experience should be compelling in the sense that the users should be more concerned with their own actions and behaviors than how the interactive system works. Finally, it was desired that the system not break suspension of disbelief by doing something either inappropriate (e.g. encouraging users to stand still even though they are not moving) or overtly computer-like (repeating identical warnings or admonishments many times).

The immersive power of a compelling storyline cannot be overstated when constructing a space 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-crafted story was used to make participants more likely to suspend disbelief and more curious and less apprehensive about what will happen next. The story ties the physical space, the participant's actions, and the different output media together into a coherent, rich, and therefore immersive experience.

An immersive space is most engaging when participants believe their actions are having an effect upon the environment by influencing the story. The KidsRoom uses computer vision to achieve this goal by making the room responsive to the position, movements, and actions of the children. The room responds when children are near particular pieces of furniture. The room transforms the instant that everyone

jumps on the bed, and the kids can move and steer their "magic boat" using natural body gestures. Monsters follow the dance moves of the children,; monsters growling stop when children hide behind the bed. Monsters react when children scream loudly, but the narrator reacts when they don't. Immediately upon their first interaction with the room the children realize that what they do makes a difference in how the room responds. This perceptual sensing enhances and energizes the narrative.

A goal that was critical to obtaining the 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 question why they need to engage in some particular behavior.

No matter how well an interactive storyline 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 degrades gracefully.

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

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, but if the person continues to cause trouble, the system will continue to progress through the story anyway.

Finally, to give the KidsRoom narrative a cohesive, immersive feel, there are thematic threads that run throughout the story. For example, stuffed animals on the walls in the children's bedroom are similar to the monsters that appear later in the story. 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.

Children as subjects

Davenport and Friedlander (Davenport & Friedlander 1995) and Druin and Perlin (Druin & Perlin 1994) 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. Children already like to play with each other in real spaces engaging in rigorous physical activity, and we thought that they would be more likely than adults to be motivated by supplementary imagery, sound effects, music, and lighting.

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

Children did provide unique challenges as well. The behavior of children, particularly their group behavior, is difficult to predict. Further, children have short attention spans and often move about with explosive energy, leaving the longer playing narrations behind. Young children are small compared to adults, which can create problems for developing vision algorithms. Having children as subjects made it especially difficult to iteratively develop and test each component of the system on the primary users. Kids would tire quickly of any sort of repetitive testing of a particular component of the room. But, there is no question that the magic of the children's play is what made the magic of the KidsRoom come to life.

Visual and audio feedback

To further achieve the goal of the keeping the action in the real, as opposed to virtual, space, we designed the visual and audio feedback to only minimally focus the attention of the children. Typically, virtual reality systems use semi-realistic three-dimensional rendered scenes and video as the primary form of system feedback. We decided, however, that in order to give the room a magical, theatrical feel and in order to keep the emphasis of the space in the room and not on the screens, images would have a two-dimensional 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 mood-setting backdrops and are not always the center of the participants' attention.

Audio is the main form of feedback in the room and has several advantages over video. Realistic or semi-realistic three-dimensional video animation, even of cartoon monsters, is difficult to do well. It is much easier to obtain rich and realistic sound effects and narration. More 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. 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 that leaves much 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.

Physical constraints

A major challenge when designing the KidsRoom was to minimize the impact of our physical constraints on the development of an interesting story. This section expands on a few constraints that impact the design of perceptually-based immersive environments.

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(see (Bobick *et al.* 1996) for a review). 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.

Projection screens are light sources

One serious obstacle 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.

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 (Cruz-Neira, Sandin, & DeFanti 1993), 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 (Ivanov, Bobick, & Liu 1998).

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 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 trade-off 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. Further, 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. In fact, keeping the rugs away from the screen, away from each other, and within the camera view meant there was only one configuration in which they could be placed. The bed and the back two rugs were also under tight positional constraints.

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 so its image would not change.

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.

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 provide reasonable localization when listeners are near the center of the room. However, when a participant is near a loudspeaker playing a sound, that single speaker tends to dominate the positional percept and the spatial illusion breaks down. Sound localization is important if real objects in a space are to be given "personalities" using sound effects, 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 speech recognition. Unlike some other immersive environments that require portable microphones (Coen 1997), 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.

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. We present several in an effort to prevent others from repeating our mistakes.

Group vs. individual activity

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

In general, the more children 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. Also, a lone child often requires more intervention from the system to guide him or her through the experience.

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

Exploratory vs. linear spaces

In our initial design of the KidsRoom, we had planned to create a mostly exploratory space, modeled somewhat on popular non-linear computer games like Myst (Broderbund Software 1994). 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. For example, stepping on a rug with a frog on it would make a "ribbit" sound and moving near a real or virtual window and putting an arm towards the window generated a loud crash of glass.

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. 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 (less than 500 ms) lag in the perceptual system. 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 similar problems with the second (forest) world.

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 (Druin & Perlin 1994; Davenport & Friedlander 1995). 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.

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.

Avoiding repetitiveness

One way to break the immersive experience is for the system to repeat a single narration 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 are not heard or are ignored. For example, in the dance segment of the monster world, the control program continually checks if someone has stepped off their rug or if two people are on the same rug. If someone drifts off a rug more than once, narration is needed, but repeating a narration just played moments before imparts a mechanical sense to the responses and causes the entire experience to feel less 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.

Sensitivity to timing

People are sensitive to small timing problems and the control architecture proved to be sensitive to small changes in timing values. 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 of a four second duration begins to play. During the first second, someone gets off the bed. However, the man-overboard narration (which lasts four seconds as well) cannot play because the first narration has not yet completed, and the system does not 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 eventually plays but three 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.

Another difficulty encountered was the tuning of timer parameters used by the control program (see (Bobick *et al.* 1996) for a discussion of the control architecture). 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 (Pinhanez & Bobick 1997) 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. 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.

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 the room explicitly indicating that it knows someone is still off the bed, and an opportunity to demonstrate responsiveness is wasted.

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 system 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 cannot recognize (i.e. the allowed dance moves in the monster land).

Summary and contributions

The KidsRoom went from whiteboard sketches to a full installation in eight weeks. 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.

We believe the KidsRoom provides several fundamental contributions. First, is the demonstration that non-encumbering sensors can be used for the measurement and recognition of individual and group action in a complex interactive narrative. Second, unlike most previous interactive systems, the KidsRoom does not require the user to wear special clothing, gloves or vests, does not require embedding sensors in objects, and has been explicitly designed to allow multiple simultaneous users. Third, the KidsRoom moves beyond just measurement of position towards recognition of action using measurement *and* context – particularly context generated by the system itself. 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.[4]

References

Bobick, A.; Intille, S.; Davis, J.; Baird, F.; Campbell, L.; Ivanov, Y.; Pinhanez, C.; Schütte, A.; and Wilson, A. 1996. The KidsRoom: A perceptually-based interactive and immersive story environment. M.I.T. Media Laboratory Perceptual Computing Section 398, M.I.T. Media Laboratory Perceptual Computing Section. Revised September 1997, see also http://vismod.www.media.mit.edu/vismod/demos/kidsroom.

Broderbund Software. 1994. Myst. An interactive CD-ROM.

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

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

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

[4] 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.

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

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

Intille, S.; Davis, J.; and Bobick, A. 1997. Real-time closed-world tracking. In *Proceedings of the IEEE Computer Society Converence on Computer Vision and Pattern Recognition*, 697–703. Los Alamitos, CA: IEEE Computer Society.

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

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

Pinhanez, C., and Bobick, A. 1997. Human action detection using PNF propagation of temporal constraints. Technical Report 423, M.I.T. Media Laboratory Perceptual Computing Section, 20 Ames Street, Cambridge, MA 02139.