# The KidsRoom: An example application using a deep perceptual interface

Aaron Bobick Jim Davis Stephen Intille
The MIT Media Laboratory
20 Ames Street
Cambridge, MA 02139
{bobick, jdavis, intille}@media.mit.edu

## 1 Overview: The *KidsRoom*

The KidsRoom is a perceptually-based interactive, narrative playspace for children constructed at the MIT Media Lab. The design of the application was motivated by the following goals: (1) to keep the focus of user action and interaction in the physical, as opposed to 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.

The space re-creates a child's bedroom and measures of 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 color theatrical lighting on the ceiling illuminates the space. A microphone is placed inside the room to detect the volume of shouts. Four speakers, one on each wall, project directional sound effects and music into the space. Finally, there are four video cameras in the space. Three are used for computer vision, to recognize the actions of the children, and one is used for spectators to view the room when people are using it. Figure 1 shows a view of the complete KidsRoom installation.

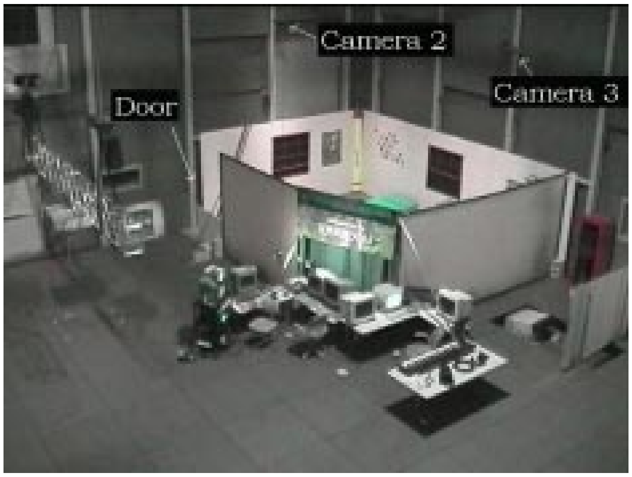

**Figure 1:** The KidsRoom is a 24 by 18 foot space constructed in our lab. Two walls resemble the walls in a real children's room, complete with posters and windows. The other two walls are large back-projection screens. Six computers used to control the fully-automated room and additional audio and 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.

The basic experience of the KidsRoom is inspired by numerous famous children's stories (such as *Peter Pan* and *Where the Wild Things Are*) in which children's rooms are transformed into magical places. The narrative carries up to four children through four scenes: (1) the bedroom, where the children must learn the magic word from the furniture, (2) forest world, where cooperative traversal along a path is enforced by hidden monsters, (3) river land, where the bed becomes a boat and paddling by the children moves the boat down a projected river, and (4) monster world, where fantastical monsters first teach the children four dance steps and then mimic the children as the they perform the learned moves. Throughout the experience the room appropriately responds to the actions of the children by changing projected video images on the two video walls, moving sound effects around the room, providing guiding narration and hints, varying the lighting, and playing individually scored musical accompaniments. The entire experience lasts between 10 and 12 minutes depending upon the behavior of the children.[1]

We believe the KidsRoom is the first multi-person, fully-automated interactive, narrative playspace ever constructed. We also believe that it is a strong example of a *deep* perceptual interface. By deep we really mean *ad hoc* in the good sense: the context of the content controls which properties are measured and how they are interpreted. The goal of this abstract is to briefly describe some of the perceptual user interface technologies employed, and how the tight coupling between the context of the interaction and the sensing mechanisms resulted in a powerful control system. More details about the design, implmentation, and evaluation of the KidsRoom can be found in [?].

## 2 Perceptual tasks

All of the perception performed in the KidsRoom was performed using computer vision with exception of the detecting the volume at which the children shouted out the magic word required to transform the bedroom. Three cameras were used: one overhead, two oblique views. In all about a dozen distinct "actions" were recognized by the system. Here we enumerate some of perceptual tasks and discuss the technologies exploited and the degree to which the methods were designed to work explicitly in the relevant contexts.

### 2.1 Tracking

Unlike previous non-wired, visually-based interactive systems (e.g. [?]) the KidsRoom was designed for multiple simultaneous users: up to four children could be in the room at the same time. This design goal resulted in the making the general perceptual task of tracking more difficult than in most such systems. Furthermore, the children frequently collide, all row on the bed at one time, and run when the monster sounds became loud and chased them behind the bed (sound was directionally controlled using four speakers). These activities complicate tracking further.

To accurately track the people we implemented a form of *closed-world tracking* for use on imagery from the overhead camera; the details of the algorithm are given in [?]. The most significant aspect is that the tracking strategy was designed based upon the context of the environment: the presence of a door, the color of a child did not change dramatically over the course of the action, the image regions of the children could be separated from the background. Figure 2 illustrates the overhead imagery used for tracking and the blob representation used to maintain identity.

The tracker did make some errors, but almost always errors of identity, not location. This behavior impacted the design of the interface: the room would assume it knew where the children were, but not necessarily which child was which.

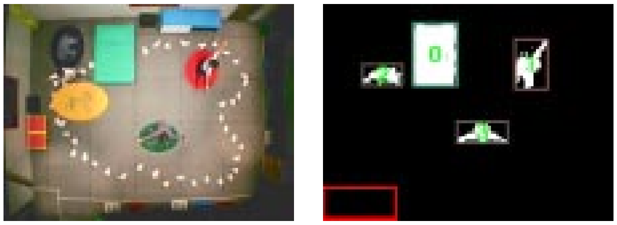

**Figure 2:** The left image shows a view with three people in the room from the overhead camera used for tracking.The right image shows the output of the tracking system, which is described and evaluated elsewhere[?]. All three people and the bed are being tracked as they move about. The box in the 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.

### 2.2 Rowing

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.

The "man-overboard" detection uses the generic tracking results mentioned above. But the rowing is detected by measuring motion-energy (computed from image differencing) on each side of the bed. This information is then used by the control program to decide how fast the passengers are "rowing" and if the children have avoided obstacles in the river by rowing vigorously on the correct side of the boat.

Of course, the motion information can be interpreted as rowing only within a specific context established by the narrative. In fact, the algorithm requires that everyone is inside the boat, i.e. all on the bed. This information is determined using the tracker once the participants have received the appropriate narration. Knowing the size of the bed from the top down tracker background-subtracted view, the algorithm waits until the number of pixels in the large blob 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 everyone is "in the boat" the blob size is about right and the bed orientation (i.e. left and right side of bed) is computed. Example imagery and estimation of rowing motion is shown in Figure 3.

[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.

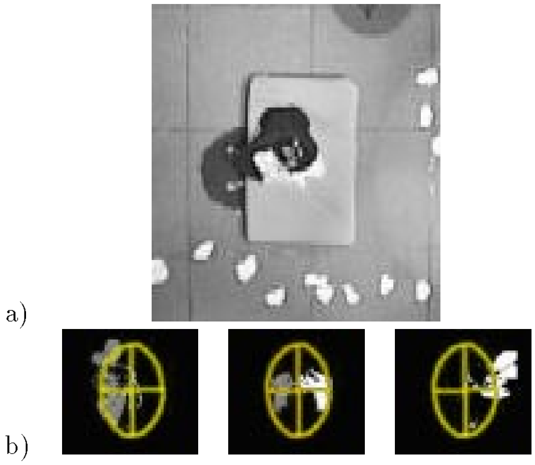

**Figure 3:** These images show the motion energy that is detected from the overhead camera (a) when a person is "rowing" as they sit on top of the bed. (b) The ellipse represents the position and orientation of the bed, extracted by the system. The colored pixels indicate where the system detected motion. 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.

Again, the user interface is deeply connected to the content. We did not define a generic rowing detector. We constructed an ad hoc method of detecting the necessary information in this context. Importantly, the context was enforced by the environment: if a person climbed off the bed, a loud and urgent "man-overboard" narration quickly forced the errant rower back on. The result was an extremely robust interaction.

### 2.3 Monster dancing

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 carefully 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.

The four different moves were recognized using different mechanisms. Briefly, the crouch was seen by looking for a temporal signature in variation of height; the 'Y' by a static moment computation; the flap and spin by *temporal templates* [?]. Figures 4 and 5 illustrate the moves and the methods. Because we used independent detectors it was possible that more than one move would be detected at the same time. Here we again used the fact that we had complete knowledge of context to disambiguate between likely confused pairs. For example since a 'Y' is contained in the flap, we used a simple temporal analysis to disambiguate between the two moves if they were both detected simultaneously.

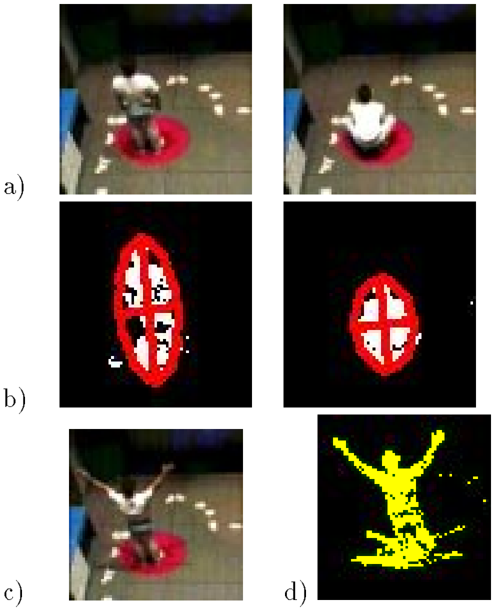

**Figure 4:** (a) A person performing a crouch move. (b) A person's background difference blob. Overlayed on top is an ellipse model of the blob. The first image shows a person in the standing position. The second shows the same person crouching. The difference in elongation of the ellipse model is used 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.

## 3 Conclusion

The KidsRoom went from whiteboard sketches to a full installation in eight weeks. The time constraints prevented us from exploring other visually-based perceptual inference issues such as reasoning about occlusion in a multi-person environment (we intended to use the overhead view to select oblique views for each person). But the experience demonstrated the power of tightly coupling perceptual user interfaces with the semantics and context of the content. On the primary day of the exhibit we had over 100 children play in the KidsRoom and it worked every time. That degree of robustness requires deep connections between interface and content.

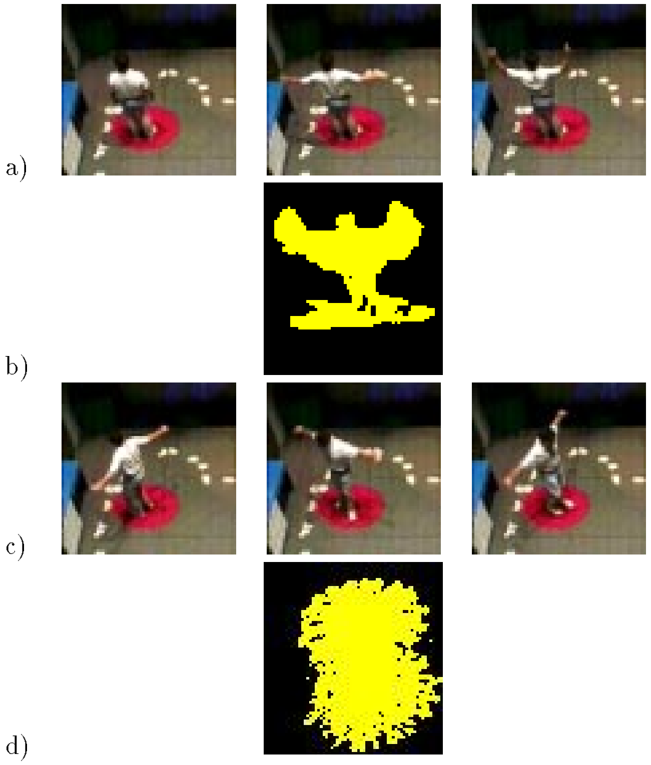

**Figure 5:** Two of the dance move actions are recognized using a motion template matching method developed by Davis and Bobick[?]. (a), (c) A person doing a flap move and a spin. The system detects the flaps and spins by matching motion models (which have been computing using a database of example moves) to the motion templates shown (b), (d). Note in the flap example the top part of the blob is generated by the moving arms while the bottom part is generated by shadows from the arms. In the KidsRoom, shadows were incorporated into the models of the moves.

## References

[1] A. Bobick, S. Intille, J. Davis, F. Baird, L. Campbell, Y. Ivanov, C. Pinhanez, A. Schutte, and A. Wilson. 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, November 1996. Revised September 1997, see also http://vismod.www.media.mit.edu/vismod/demos/kidsroom.

[2] J.D. Davis and A.F. Bobick. The representation and recognition of action using temporal templates. In *Proc. Comp. Vis. and Pattern Rec.*, pages 928–934. IEEE Computer Society Press, June 1997.

[3] 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. IEEE Computer Society Press, June 1997.

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