Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)

Spatial Scaffolding for Sociable Robot Learning

Cynthia Breazeal, Matt Berlin
MIT Media Lab
20 Ames St., Cambridge, MA 02139
{cynthiab, mattb}@media.mit.edu

Abstract

Spatial scaffolding is a naturally occurring human teaching behavior, in which teachers use their bodies to spatially structure the learning environment to direct the attention of the learner. Robotic systems can take advantage of simple, highly reliable spatial scaffolding cues to learn from human teachers. We present an integrated robotic architecture that combines social attention and machine learning components to learn tasks effectively from natural spatial scaffolding interactions with human teachers. We evaluate the performance of this architecture in comparison to human learning data drawn from a novel study of the use of embodied cues in human task learning and teaching behavior. This evaluation provides quantitative evidence for the utility of spatial scaffolding to learning systems. In addition, this evaluation supported the construction of a novel, interactive demonstration of a humanoid robot taking advantage of spatial scaffolding cues to learn from natural human teaching behavior.

Introduction

How can we design robots that are competent, sensible learners? Learning will be an important part of bringing robots into the social, cooperative environments of our workplaces and homes. Our research seeks to identify simple, non-verbal cues that human teachers naturally provide that are useful for directing the attention of robot learners. The structure of social behavior and interaction engenders what we term "social filters:" dynamic, embodied cues through which the teacher can guide the behavior of the robot by emphasizing and de-emphasizing objects in the environment.

This paper describes a novel study that we conducted to examine the use of social filters in human task learning and teaching behavior. Through this study, we observed a number of salient attention-direction cues. In particular, we argue that spatial scaffolding, in which teachers use their bodies to spatially structure the learning environment to direct the attention of the learner, is a highly valuable cue for robotic learning systems.

In order to directly evaluate the utility of the identified cues, we integrated novel social attention and learning mechanisms into a large architecture for robot cognition. Working together, these mechanisms take advantage of the structure of nonverbal human teaching behavior, allowing the robot to learn from natural spatial scaffolding interactions. We evaluated the performance of this integrated learning architecture in comparison to human learning data on benchmark tasks drawn from the study, providing quantitative evidence for the utility of the identified cues. Additionally, we constructed a novel, interactive demonstration of a humanoid robot learning tasks from natural human teaching behavior.

There has been a large, interesting body of work focusing on human gesture, especially communicative gestures closely related to speech (Cassell 2000; McNeill 1992). In the computer vision community, there has been significant prior work on technical methods for tracking head pose (Morency et al. 2002) and for recognizing hand gestures such as pointing (Wilson and Bobick 1999). Others have contributed work on using these cues as inputs to multimodal interfaces (Bolt 1980). Such interfaces often specify fixed sets of gestures for controlling systems such as graphical expert systems (Kobsa et al. 1986), natural language systems (Neal et al. 1998), and even directable robotic assistants (Fransen et al. 2007).

However, despite a large body of work on understanding eye gaze (Langton 2000), much less work has been done on using other embodied cues to infer a human's emphasis and de-emphasis in behaviorally realistic scenarios. One of the important contributions of this work is the analysis of spatial scaffolding cues in a human teaching and learning interaction, and the empirical demonstration of the utility of spatial scaffolding for robotic learning systems. In particular, our work identifies a simple, reliable, component of spatial scaffolding: attention direction through object movements towards and away from the body of the learner.

Emphasis Cues Study

A set of tasks was designed to examine how teachers emphasize and de-emphasize objects in a learning environment with their bodies, and how this emphasis and de-emphasis guides the exploration of a learner and ultimately the learning that occurs.

We gathered data from 72 individual participants, combined into 36 pairs. For each pair, one participant was randomly assigned to play the role of teacher and the other par-

ticipant assigned the role of learner for the duration of the study. For all of the tasks, participants were asked not to talk, but were told that they could communicate in any way that they wanted other than speech. Tasks were presented in a randomized order.

For all of the tasks, the teacher and learner stood on opposite sides of a tall table, with 24 colorful foam building blocks arranged between them on the tabletop. These 24 blocks were made up of four different colors - red, green, blue, and yellow, with six different shapes in each color - triangle, square, small circle, short rectangle, long rectangle, and a large, arch-shaped block.

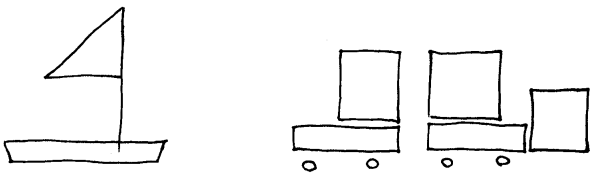

Figure 1: Task instruction cards given to learners.

The two study tasks were interactive, secret constraint tasks, where one person (the learner) knows what the task is but does not know the secret constraint. The other person (the teacher) doesn't know what the task is but does know the constraint. So, both people must work together to successfully complete the task. For each of the tasks, the learner received instructions, shown in figure 1, for a figure to construct using the blocks. In Task 1, the learner was instructed to construct a sailboat figure using at least 7 blocks; in Task 2, a truck/train figure using at least 8 blocks. When put together with the secret constraints, the block number requirements turned these tasks into modestly difficult Tangram-style spatial puzzles.

The secret constraint handed to the teacher for Task 1 was that "the figure must be constructed using only blue and red blocks, and no other blocks." The secret constraint for Task 2 was that "the figure must include all of the triangular blocks, and none of the square blocks." At the end of each task, the learner was asked to write down what they thought the secret constraint might have been.

Study Observations

Since neither participant had enough information to complete the task on their own, these tasks required the direct engagement and cooperation of both participants. Correspondingly, we observed a rich range of dynamic, interactive behaviors during these tasks.

To identify the emphasis and de-emphasis cues provided by the teachers in these tasks, an important piece of "ground-truth" information was exploited: for these tasks, some of the blocks were "good," and others of the blocks were "bad." In order to successfully complete the task, the teacher needed to encourage the learner to use some of the blocks in the construction of the figure, and to steer clear of some of the other blocks. For example, in Task 1, the blue and red blocks were "good," while the green and yellow blocks were "bad."

We observed a wide range of embodied cues provided by the teachers in the interactions for these two tasks, as well as a range of different teaching styles. Positive emphasis cues included simple hand gestures such as tapping, touching, and pointing at blocks with the index finger. These cues were often accompanied by gaze targeting, or looking back and forth between the learner and the target blocks. Other positive gestures included head nodding, the "thumbs up" gesture, and even shrugging. Teachers nodded in accompaniment to their own pointing gestures, and also in response to actions taken by the learners.

Negative cues included covering up blocks, holding blocks in place, or maintaining prolonged contact despite the proximity of the learner's hands. Teachers would occasionally interrupt reaching motions directly by blocking the trajectory of the motion or even by touching or (rarely) lightly slapping the learner's hand. Other negative gestures included head shaking, finger or hand wagging, or the "thumbs down" gesture.

An important set of cues were cues related to block movement and the use of space. To positively emphasize blocks, teachers would move them towards the learner's body or hands, towards the center of the table, or align them along the edge of the table closest to the learner. Conversely, to negatively emphasize blocks, teachers would move them away from the learner, away from the center of the table, or line them up along the edge of the table closest to themselves. Teachers often devoted significant attention to clustering the blocks on the table, spatially grouping the bad blocks with other bad blocks and the good blocks with other good blocks. These spatial scaffolding cues were some of the most prevalent cues in the observed interactions. Our next step was to establish how reliable and consistent these cues were in the recorded data set, and most importantly, how useful these cues were for robotic learners.

Data Analysis

In order to record high-resolution data about the study interactions, we developed a data-gathering system which incorporated multiple, synchronized streams of information about the study participants and their environment. For all of the tasks, we tracked the positions and orientations of the heads and hands of both participants, recorded video of both participants, and tracked all of the objects with which the participants interacted.

Our data analysis pipeline is shown in figure 2. Images of the foam blocks on the table surface (1) were provided by a camera system mounted underneath the table. Color segmenation (2) was used to identify pixels that were associated with the red, green, blue, and yellow blocks, and a blob finding algorithm identified the locations of possible blocks within the segmented images. Next, a shape recognition system (3) classified each blob as one of the six possible block shapes, and an object tracking algorithm updated the positions and orientations of each block using these new observations.

To track the head and hand movements of the study participants, we employed a 10-camera motion capture system along with customized software for tracking rigid ob-

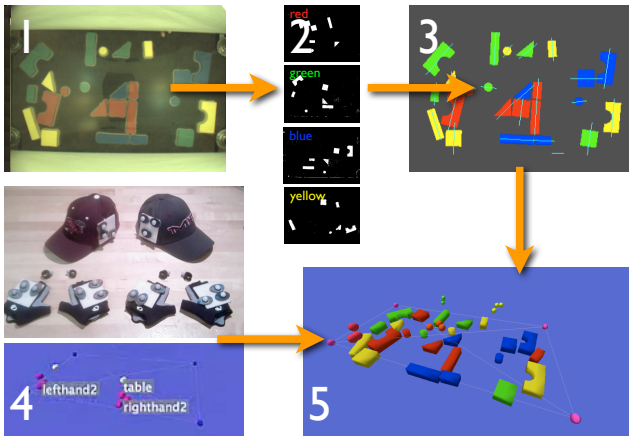

Figure 2: The study data processing pipeline.

jects (4). Study participants wore special gloves and baseball caps mounted with small, retroreflective markers that were tracked by this system. Finally, the tracking information about the foam blocks was mapped into the coordinate system of the motion capture system, so that all of the tracked study objects could be analyzed in the same, three-dimensional frame of reference (5).

With all of the study objects now in the same frame of reference, the next stage of analysis used spatial and temporal relationships between the blocks and the bodies of the participants to extract a stream of potentially salient events that occurred during the interactions. These events included, among other things, block movements and hand-to-block contact events, which were important focal points for our analysis. Our processing system recognized these events, and attempted to ascribe agency to each one (i.e., which agent - learner or teacher - was responsible for this event?). Finally, statistics were compiled looking at different features of these events, and assessing their relative utility at differentiating the "good" blocks from the "bad" blocks.

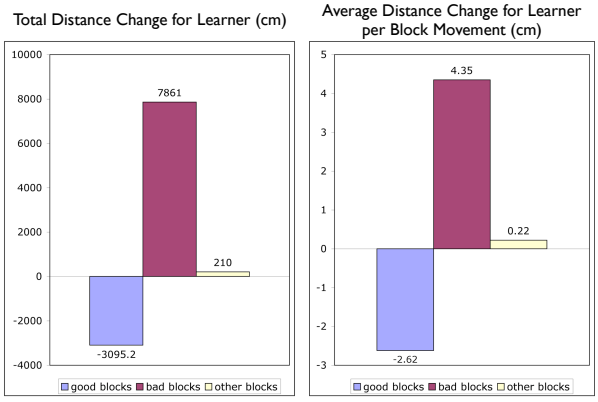

Figure 3: Change in distance to the body of the learner for block movements initiated by the teacher. Negative values represent movement towards the learner, while positive values represent movement away from the learner.

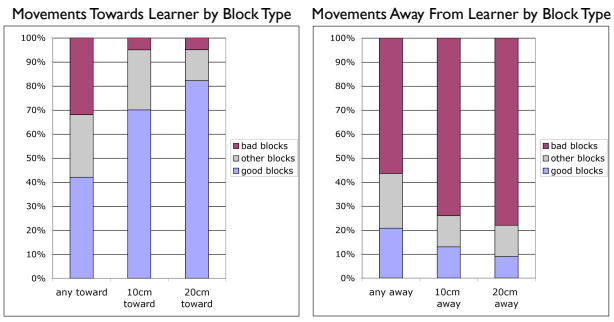

Figure 4: Movements towards the body of the learner initiated by the teacher were predictive of good blocks. Movements away from the body of the learner were predictive of bad blocks. The differentiating power of these movements increased for more substantial changes in distance towards and away.

One of the most interesting features that we analyzed was movement towards and away from the bodies of the participants. The results of our analysis are summarized in figures 3 and 4. As can be seen in figure 3, the aggregate movement of good blocks by teachers is biased very substantially in the direction of the learners, while the aggregate movement of bad blocks by teachers is biased away from the learners. In fact, over the course of all of the 72 analyzed interactions, teachers differentiated the good and bad blocks by more than the length of a football field in terms of their movements relative to the bodies of the learners.

Movements towards the body of the student were correlated with good blocks, with stronger correlations for movements that more significantly changed the distance to the learner. For changes in distance of 20cm or greater, fully 83% of such movements were applied to good blocks versus 13% for other blocks and 5% for bad blocks. A similar pattern was seen for block movements away from the body of the learner, with larger changes in distance being strongly correlated with a block being bad, as shown in figure 4.

Thus, we have identified an embodied cue which might be of significant value to a robotic system learning in this task domain. A robot, observing a block movement performed by a teacher, might be able to make a highly reliable guess as to whether the target block should or should not be used by measuring the direction and distance of the movement. Such a cue can be interpreted simply and reliably even within the context of a chaotic and fast-paced interaction.

Integrated Learning Architecture

In order to evaluate the utility of the spatial scaffolding cues identified in the study, we integrated novel social attention and learning mechanisms into a large architecture for robot cognition. Working together, these mechanisms take advantage of the structure of nonverbal human teaching behavior, allowing the robot to learn from natural spatial scaffolding interactions. Our implementation enabled the evaluation of our architecture's performance on benchmark tasks drawn from the studies, and also supported the creation of an interactive, social learning demonstration.

Our integrated learning architecture incorporates simulation-theoretic mechanisms as a foundational and organizational principal to support collaborative forms of human-robot interaction. An overview of the architecture, based on (Gray et al. 2005), is shown in Figure 5. Our implementation enables a humanoid robot to monitor an adjacent human teacher by simulating his or her behavior within the robot's own generative mechanisms on the motor, goal-directed action, and perceptual-belief levels.

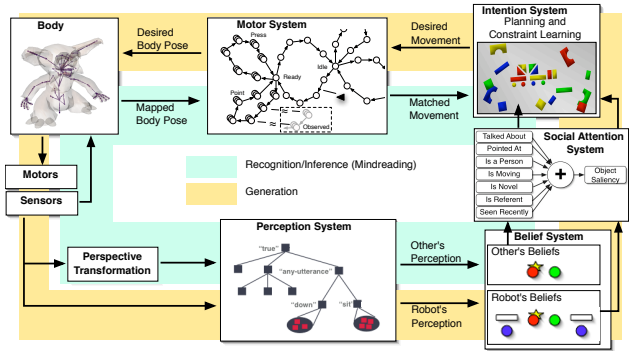

Figure 5: System architecture overview.

Social Attention Mechanisms

The mechanisms of social attention integrated into our cognitive architecture help to guide the robot's gaze behavior, action selection, and learning. These mechanisms also help the robot to determine which objects in the environment the teacher's communicative behaviors are *about*.

Shared attention is a critical component for human-robot interaction. Gaze direction in general is an important, persistent communication device, verifying for the human partner what the robot is attending to. Additionally, the ability to share attention with a partner is a key component to social attention (Scassellati 2001).

Referential looking is essentially "looking where someone else is looking". Shared attention, on the other hand, involves representing mental states of self and other (Baron-Cohen 1991). To implement shared attention, the system models both the attentional focus (what is being looked at right now) and the referential focus (the shared focus that activity is *about*). The system tracks the robot's attentional focus, the human's attentional focus, and the referential focus shared by the two.

The robot's attentional system computes the saliency (a measure of interest) for objects in the perceivable space. Overall saliency is a weighted sum of perceptual properties (proximity, color, motion, etc.), the internal state of the robot (i.e., novelty, a search target, or other goals), and social cues (if something is pointed to, looked at, talked about, or is the referential focus saliency increases). The item with the highest saliency becomes the current attentional focus of the robot, and determines the robot's gaze direction.

The human's attentional focus is determined by what he or she is currently looking at. Assuming that the person's head orientation is a good estimate of their gaze direction, the robot follows this gaze direction to determine which (if any) object is the attentional focus.

The mechanism by which infants track the referential focus of communication is still an open question, but a number of sources indicate that looking time is a key factor. This is discussed in studies of word learning (Baldwin and Moses 1994; Bloom 2002). For example, when a child is playing with one object and they hear an adult say "It's a modi", they do not attach the label to the object they happen to be looking at, but rather redirect their attention to look at what the adult is looking at, and attach the label to this object.

For the referential focus, the system tracks a $relative - looking - time$ for each of the objects in the robot's environment (relative time the object has been the attentional focus of either the human or the robot). The object with the most $relative - looking - time$ is identified as the referent of the communication between the human and the robot.

Constraint Learning and Planning Mechanisms

In order to give the robot the ability to learn from embodied, spatial scaffolding cues in the secret-constraint task domain of our study tasks, we developed a simple, Bayesian learning algorithm. The learning algorithm maintained a set of classification functions which tracked the relative odds that the various block attributes were good or bad according to the teacher's secret constraints. In total, ten separate classification functions were used, one for each of the four possible block colors and six possible block shapes.

Each time the robot observed a salient teaching cue, these classification functions were updated using the posterior probabilities identified through the study - the odds of the target block being good or bad given the observed cue. At the end of each interaction, the robot identified the single block attribute with the most significant good/bad probability disparity. If this attribute was a color attribute, the secret constraint was classified as a *color* constraint. If it was a shape attribute, the constraint was classified as a *shape* constraint. Next, all of the block attributes associated with the classified constraint type were ranked from "most good" to "most bad." The learning algorithm proceeded as follows:

1: **for** each observed cue c applied to block b **do**
2: **for** each attribute a_i of block b, $a_i \in a_1, ..., a_n$ **do**
3: $P(a_i \text{ is good/bad}) * = P(b \text{ is good/bad} | c)$
4: **end for**
5: renormalize attribute distributions
6: **end for**
7: find attribute a_s where $P(a_s \text{ is good}) / P(a_s \text{ is bad})$ is most significant
8: sort all attributes a_j, $type(a_j) = type(a_s)$, based on $P(a_j \text{ is good})$

It should be noted that this learning algorithm imposed significant structural constraints on the types of rules that the robot could learn from the interactions. However, the space of rules that the robot considered was still large enough to present a significant learning challenge for the robot, with low chance performance levels. Most importantly, this learning problem was hard enough to represent an interesting evaluation of the usefulness of the identified spatial scaf-

folding cues. The core question was: would these teaching cues be sufficient to support successful learning?

A simple figure planning algorithm was developed to enable the robot to demonstrate its learning abilities. The planning algorithm allowed the robot to use a simple spatial grammar to construct target figures in different ways, allowing for flexibility in the shapes as well as the colors of the blocks used in the figures. The spatial grammar was essentially a spatially-augmented context-free grammar. Each rule in the grammar specified how a particular figure region could be constructed using different arrangements of one or more blocks. This approach allowed the robot to be quite flexibly guided by a human teacher's behavior.

For each rule in the grammar, a preference distribution specified an initial bias about which alternatives the robot should prefer. During teaching interactions, the figure planning algorithm multiplied these distributions by the estimated probability of each available block being a good block, as inferred from the teacher's embodied cues. The resulting biased probability distribution governed the robot's choice of which block to use at each step in constructing the figure.

Emphasis Cues Benchmarks

The tasks from the study were used to evaluate the ability of our integrated architecture to learn from cues that human teachers naturally provide. The robot was presented with the recorded behavioral data from the study tasks, and its learning performance was measured. After each observed task, the robot was simulated "re-performing" the given task in a non-interactive setting. The robot followed the rules extracted by its learning algorithm, and its performance was gauged as correct or incorrect according to the teacher's secret constraint.

A cross-validation methodology was followed for both of the benchmark tasks. The robot's learning algorithm was developed and tested using 6 of the 36 study sessions. The robot's learning performance was then evaluated on the remaining 30 study sessions, with 30 recorded interactions for Task 1 and 30 recorded interactions for Task 2.

The performance of the human learners and the robot on the benchmark tasks is presented in tables 1 and 2, respectively. Human performance was gauged based on the guesses that the learners wrote down at the end of each task about the secret constraint. For both tasks, the secret constraint involved two rules. The performance of the learners was gauged using three metrics: whether or not they correctly classified the rules as being color-based or shape-based (Rule Type Correct), whether or not they correctly specified either of the two rules, and finally, whether or not they correctly specified both rules. Additionally, table 2 presents how often the robot correctly re-performed the task following its observation of the interaction.

The results suggest that the robot was able to learn quite successfully by paying attention to a few simple cues extracted from the teacher's observed behavior. This is an exciting validation both of the robot's learning mechanisms as well as of the usefulness of the cues themselves. These dynamic, embodied cues are not just reliable at predicting whether blocks are good and bad in isolation. They are prevalent enough and consistent enough throughout the observed interactions to support successful learning.

Table 1: Performance of the human learners on study tasks.

Task	Rule Type Correct (color / shape)	One Rule Correct	Both Rules Correct
Sailboat	30 (100%)	27 (90%)	26 (87%)
Truck	29 (97%)	26 (87%)	4 (13%)

Table 2: Robot's learning performance on benchmark tasks.

Task	Rule Type	One Rule	Both Rules	Correct Perform.
Sailboat	24 (80%)	22 (73%)	21 (70%)	21 (70%)
Truck	28 (93%)	23 (77%)	14 (47%)	23 (77%)

Interactive Demonstration and Evaluation

Finally, we created a demonstration which featured our robot making use of spatial scaffolding to learn from live interactions with human teachers, in a similar, secret-constraint task domain. A mixed-reality workspace was created so that the robot and the human teacher could both interact gesturally with animated foam blocks on a virtual tabletop.

The teacher's head and hands were tracked using the same motion-capture tracking pipeline employed in the study. The human manipulated the virtual blocks via a custom-built gestural interface, which essentially converted an upturned plasma display into a very large, augmented touch screen (see figure 6). The interface allowed the teacher to use both hands to pick up, slide, and rotate the blocks on the screen.

Figure 6 shows off an interaction sequence between the robot and a human teacher. (1) The robot, instructed to build a sailboat figure, starts to construct the figure as the teacher watches. The teacher's goal is to guide the robot into using only blue and red blocks to construct the figure. (2) As the interaction proceeds, the robot tries to add a green rectangle to the figure. The teacher interrupts, pulling the block away from the robot. (3) As the robot continues to build the figure, the teacher tries to help by sliding a blue block and a red block close to the robot's side of the screen. (4) The teacher then watches as the robot completes (5) the figure successfully. (6) To demonstrate that the robot has indeed learned the constraints, the teacher walks away, and instructs the robot to build a new figure. Without any intervention from the teacher, the robot successfully constructs the figure, a smiley-face, using only red and blue blocks.

To evaluate the effectiveness of this interaction, we conducted a small user study. We gathered data from 18 new participants, with two task interactions per participant, for a total of 36 task interactions. The identical task protocol was followed as was used in the human-human study, with the robot playing the role of the learner. The human teach-

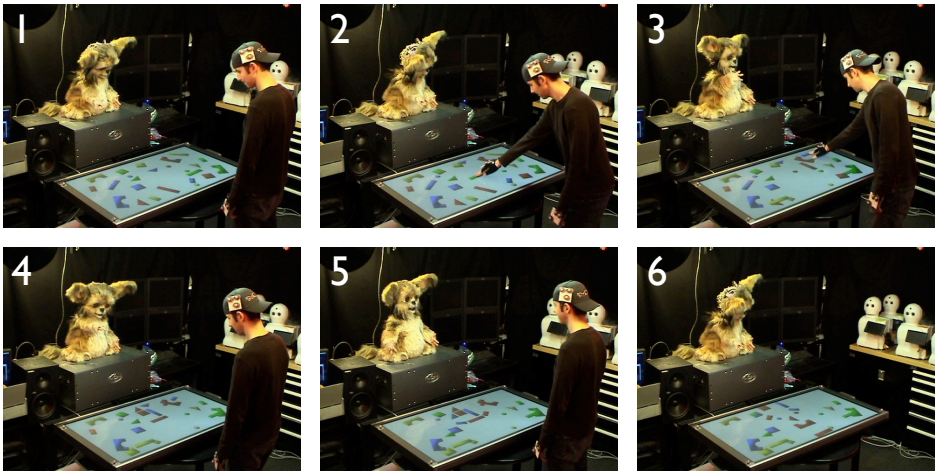

Figure 6: Interaction sequence between the robot and a human teacher.

ers were given no prompting as to what cues or behaviors the robot would be attending to. The robot was able to successfully complete the task, obeying the secret constraint, in 33 of the 36 interactions (92%). These results support the conclusion that the spatial scaffolding cues observed in human-human teaching interactions do indeed transfer over into human-robot interactions, and can be effectively taken advantage of by our integrated learning architecture.

Conclusion

This paper makes the following contributions. First, we presented a novel study of the use of embodied cues in human task learning and teaching behavior. Through this study, we identified a number of simple, highly reliable spatial scaffolding cues that robotic systems can use to learn from human teachers. Second, we presented an integrated learning architecture that combines social attention and machine learning components to learn tasks effectively from nonverbal interactions with human teachers. Finally, we evaluated the performance of this architecture in comparison to human learning data drawn from our study, and presented an interactive demonstration of a humanoid robot taking advantage of spatial scaffolding cues to learn from natural human teaching behavior. Spatial scaffolding, in which teachers use their bodies to spatially structure the learning environment to direct the attention of the learner, is a highly valuable source of information for interactive learning systems.

Acknowledgments

The work presented in this paper is a result of ongoing efforts of the graduate and undergraduate students of the MIT Media Lab Personal Robots Group. This work is funded by the *Digital Life* and *Things That Think* consortia of the MIT Media Lab. The authors would like to thank Toyota Motor Corporation, Partner Robot Division for their support.

References

Baldwin, D., and Moses, J. 1994. Early understanding of referential intent and attentional focus: Evidence from language and emotion. In Lewis, C., and Mitchell, P., eds., *Children's Early Understanding of Mind*. New York: Lawrence Erlbaum Assoc.

Baron-Cohen, S. 1991. Precursors to a theory of mind: Understanding attention in others. In Whiten, A., ed., *Natural Theories of Mind*. Oxford, UK: Blackwell Press. 233–250.

Bloom, P. 2002. Mindreading, communication and the learning of names for things. *Mind and Language* 17(1 and 2):37–54.

Bolt, R. 1980. "Put-that-there": Voice and gesture at the graphics interface. *Proceedings of the 7th annual conference on Computer graphics and interactive techniques* 262–270.

Cassell, J. 2000. *Embodied Conversational Agents*. MIT Press.

Fransen, B.; Morariu, V.; Martinson, E.; Blisard, S.; Marge, M.; Thomas, S.; Schultz, A.; and Perzanowski, D. 2007. Using vision, acoustics, and natural language for disambiguation. In *Proceedings of the 2007 ACM Conference on Human-Robot Interaction*.

Gray, J.; Breazeal, C.; Berlin, M.; Brooks, A.; and Lieberman, J. 2005. Action parsing and goal inference using self as simulator. In *14th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN)*. Nashville, Tennessee: IEEE.

Kobsa, A.; Allgayer, J.; Reddig, C.; Reithinger, N.; Schmauks, D.; Harbusch, K.; and Wahlster, W. 1986. Combining deictic gestures and natural language for referent identification. In *11th International Conference on Computational Linguistics*.

Langton, S. 2000. The mutual influence of gaze and head orientation in the analysis of social attention direction. *The Quarterly Journal of Experimental Psychology: Section A* 53(3):825–845.

McNeill, D. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Morency, L.-P.; Rahimi, A.; Checka, N.; and Darrell, T. 2002. Fast stereo-based head tracking for interactive environment. In *Int. Conference on Automatic Face and Gesture Recognition*.

Neal, J.; Thielman, C.; Dobes, Z.; Haller, S.; and Shapiro, S. 1998. Natural language with integrated deictic and graphic gestures. *Readings in Intelligent User Interfaces* 38–51.

Scassellati, B. 2001. Foundations for a theory of mind for a humanoid robot. *Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, PhD Thesis*.

Wilson, A. D., and Bobick, A. F. 1999. Paramteric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(9).