2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

Spatial Scaffolding Cues for Interactive Robot Learning

Matt Berlin, Cynthia Breazeal, Crystal Chao
MIT Media Lab

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 via a human subjects experiment which examines our humanoid robot's ability to learn from live interactions with human teachers in a secret-constraint task domain. This evaluation provides quantitative evidence for the utility of spatial scaffolding cues to systems that learn from natural human teaching behavior.**

I. 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 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 via a human subjects experiment which examined our humanoid robot's ability to learn from live interactions with human teachers in a secret-constraint task domain.

II. Background and Related Literature

Inspired by the way people and animals learn from others, researchers have begun to investigate various forms of social learning and interactive training techniques, such as imitation-based learning [1], clicker training [2], learning by observation [3], and tutelage [4].

There has also been a large, interesting body of work focusing on human gesture, especially communicative gestures closely related to speech [5], [6]. In the computer vision community, there has been significant prior work on technical methods for tracking head pose [7] and for recognizing hand gestures such as pointing [8]. Others have contributed work on using these cues as inputs to multi-modal interfaces [9], [10]. Such interfaces often specify fixed sets of gestures for controlling systems such as graphical expert systems [11], natural language systems [12], and even directable robotic assistants [13].

However, despite a large body of work on understanding eye gaze [14], [15], 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.

III. 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 participant 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.

The two study tasks were interactive, secret constraint tasks, where one person (the learner) knows what the task

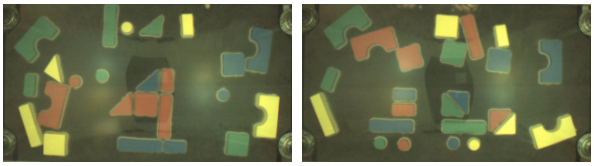

Fig. 1. Successfully completed figures (a sailboat and a truck/train with two cars) for the secret-constraint tasks. The experimental task uses 24 blocks of 6 different shapes and 4 different colors.

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, 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 (see Figure 1).

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.

A. 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."

To set the stage, we will first describe two pairs of study interactions before diving into a more detailed analysis of the observed cues. The first pair of interactions were for 1, where the goal was to construct a sailboat figure using only red and blue blocks. In one recorded interaction (session 27), the teacher is very proactive, organizing the blocks almost completely before the learner begins to assemble the figure. The teacher clusters the yellow and green blocks on one side of the table and somewhat away from the learner. The learner initially reaches for a yellow triangle. The teacher shakes her head and reaches to take the yellow block back away from the learner, before continuing to organize the blocks. The learner proceeds to complete the task successfully.

In another recorded interaction (session 7), the teacher's style is very different. Instead of arranging the blocks ahead of time, he waits for the learner to make a mistake, and then "fixes" the mistake by replacing the learner's block with one that fits the constraint. When the learner positions a green rectangle as part of the mast of the sailboat figure, the teacher quickly reaches in, pulls the block away, and replaces it with a red rectangle. Later, the teacher fixes a triangular part of the sail in a similar way, after which the learner completes the task successfully.

The second pair of interactions were for Task 2, where the goal was to construct a truck/train figure using all of the triangular blocks, and none of the square blocks. In one interaction (session 2), the teacher provides some very direct structuring of the space, pulling the square blocks away from the learner and placing the triangular blocks in front of her. In contrast, in another interaction (session 21), the teacher almost entirely refrains from moving the blocks. She instead provides gestural feedback, tapping blocks and shaking her hand "no" when the learner moves an inadmissible block, and nodding her head when the learner moves an acceptable block.

As these descriptions suggest, 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.

B. 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.

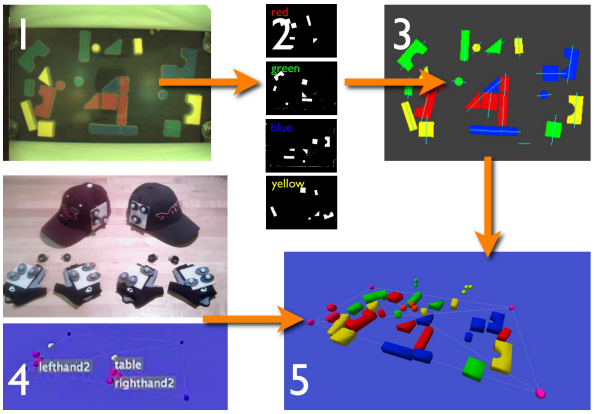

Fig. 2. The study data processing pipeline.

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 objects (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.

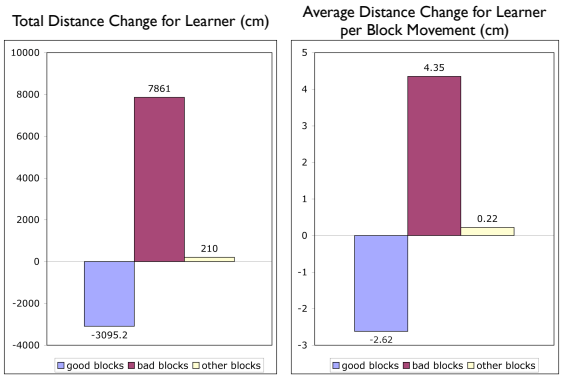

Fig. 3. Change in distance to the body of the learner for block movements initiated by the teacher. Negative values represent movement towards the learner, while positive values represent movement away from the learner.

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

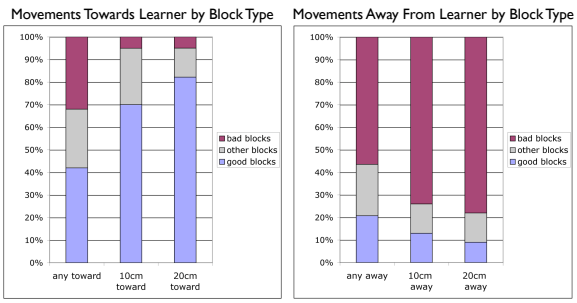

Fig. 4. Movements towards the body of the learner initiated by the teacher were predictive of good blocks. Movements away from the body of the learner were predictive of bad blocks. The differentiating power of these movements increased for more substantial changes in distance towards and away.

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.

IV. 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 creation of an interactive, social learning demonstration with our humanoid robot, and an evaluation of the robot's ability to learn from live interactions with human teachers on benchmark tasks drawn from the study.

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 [2] and [16], 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.

A. 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 [17].

Referential looking is essentially "looking where someone else is looking". Shared attention, on the other hand, involves representing mental states of self and other [18]. 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

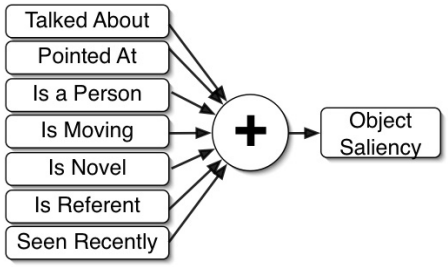

Fig. 6. Saliency of objects and people are computed from several environmental and social factors.

(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 [19], [20]. 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.

B. 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

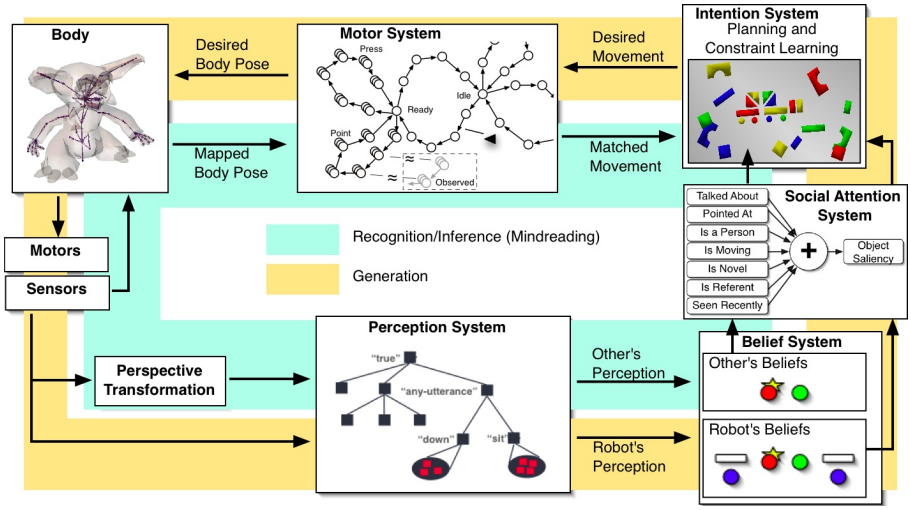

Fig. 5. System architecture overview.

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 scaffolding 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.

V. Interactive Demonstration and Evaluation

In order to evaluate the utility of spatial scaffolding cues and the effectiveness of our learning architecture, we conducted a human subjects experiment which examined our robot's ability to learn from live interactions with human teachers in a secret-constraint task domain. The subjects were not told how to teach the robot other than to avoid using verbal communication. A mixed-reality workspace using a plasma display was created so that the robot and the human teacher could both interact gesturally with animated foam blocks on a virtual tabletop, as shown in figure 9. As in the previous secret constraints study, twenty-four animated foam blocks were displayed on the screen, with 6 different block shapes and 4 different colors.

The robot could manipulate the virtual blocks directly, by procedurally sliding and rotating them on the screen. To provide feedback to the human teacher, the robot gesturally tracked these procedural movements with its hand outstretched, giving the impression that the robot was levitating the blocks with its hand. The robot also followed its own movements and those of the human with its gaze, and attended to the teacher's hands and head using an optical marker tracking system and customized object-tracking software. Additionally, the robot provided gestures of confusion and excitement at appropriate moments during the interaction (i.e., when the robot was interrupted by the human or when the figure was completed successfully).

The teacher manipulated the virtual blocks via a custom-built gestural interface, which essentially converted the up-

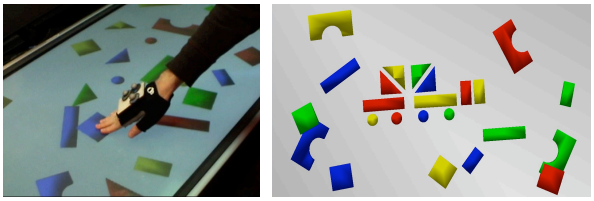

Fig. 7. A gestural interface (left) allowed the human and the robot to interact with virtual foam blocks. A figure planning algorithm allowed the robot to use a simple spatial grammar to construct figures (right) in different ways, incorporating the guidance of the teacher.

turned plasma display into a very large, augmented touch screen (see figure 7). The interface allowed the teacher to use both hands to pick up, slide, and rotate the blocks on the screen.

Replicating our previous human study, the two human-robot study tasks were interactive, secret constraint tasks. The robot learner knows the target figure to construct but does not know the secret constraint, and the human teacher knows the constraint but not the target figure. Thus, both agents must work together to successfully complete the task. For each of the tasks, the robot received instructions for a figure to construct using the blocks. As in the prior human study, the target figure was a sailboat in Task 1, and a truck/train figure in Task 2 (see Figure 1).

The human teacher received written instructions describing a constraint that the robot needed to obey in constructing the figure. 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." 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.

VI. Results

We gathered data from 18 participants, with two construction tasks per participant, for a total of 36 task interactions. The human teachers 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, learning about and 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.

In addition to measuring the robot's task performance, we also directly analyzed the recorded movement data generated by the human teachers. A similar quantitative analysis process was followed as was used in the previous human-human study. The results of the analysis are summarized in figure 8, which can be compared to the human-human data in figures 3 and 4. As before, movement of good blocks by teachers is biased in the direction of the learner (in this case the robot),

Fig. 9. The robot took advantage of spatial scaffolding cues to learn through live interactions with human teachers.

while movement of bad blocks is biased very substantially away from the robot.

Also as in the human-human study, travel towards and away from the body of the learner was strongly correlated with whether or not the given block was good or bad. Movements towards the body of the robot were correlated with good blocks, with stronger correlations for movements that more significantly changed the distance to the robot. For changes in distance of 20cm or greater, 79% of such movements were applied to good blocks versus 18% for other blocks and only 3% for bad blocks. And again, a similar pattern was seen for block movements away from the body of the robot, with larger changes in distance being strongly correlated with a block being bad. Thus, similar quantitative movement behavior was observed in our human teachers in this new study setting, which in turn supported the successful task performance of the robot in these learning interactions.

VII. Conclusion

This paper has presented the results of two studies of human nonverbal teaching behavior in an interactive, secret-constraint construction task domain. The results highlight the utility of a particular nonverbal cue - movements towards and away from the body of the learner - for indicating which objects the teacher wants the learner to interact with and which the teacher wants the learner to avoid. We argue that such spatial scaffolding behavior, 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, robotic learning systems.

Our first study recorded the teaching and learning behavior of human participants in our collaborative task domain. A range of interesting nonverbal teaching cues were observed, but our quantitative analysis highlighted the specific predictive value of movements towards and away from the body of the learner at indicating whether the target objects were good or bad.

To more directly evaluate the utility of these nonverbal

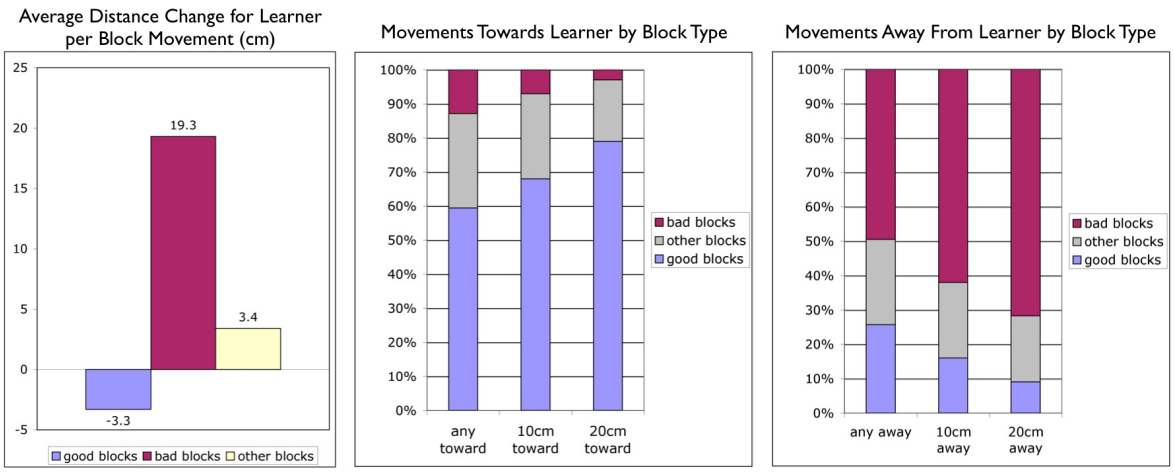

Fig. 8. Analyses of block movements towards and away from the body of the robot initiated by the human teacher. Observed movement behavior was similar to the human-human study; see figures 3 and 4 for comparison.

cues, we developed an integrated learning architecture that combines social attention and machine learning components to learn tasks from nonverbal interactions with human teachers. We evaluated the performance of this architecture via our second study: a human subjects experiment which examined our humanoid robot's ability to learn from live interactions with human teachers in a similar, secret-constraint task domain.

Our robot learned successfully from untrained human teachers in this task domain, which suggests that the spatial scaffolding behaviors observed in human-human interactions do indeed transfer over into teaching interactions with robots. Our quantitative analysis of the movements initiated by the human teachers further supports this conclusion, and highlights strong similarities between the human-human and human-robot interactive domains.

VIII. 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 generous support.

References

[1] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in Cognitive Sciences*, vol. 3, p. 233242, 1999.

[2] B. Blumberg, M. Downie, Y. Ivanov, M. Berlin, M. Johnson, and B. Tomlinson, "Integrated learning for interactive synthetic characters," in *Proceedings of the ACM SIGGRAPH*, 2002.

[3] M. M. Veloso, P. E. Rybski, and F. von Hundelshausen, "Focus: a generalized method for object discovery for robots that observe and interact with humans," in *HRI '06: Proc. of the 1st ACM conference on Human-Robot Interaction*, 2006.

[4] C. Breazeal, G. Hoffman, and A. Lockerd, "Teaching and working with robots as collaboration," in *Proceedings of the AAMAS*, 2004.

[5] J. Cassell, *Embodied Conversational Agents*. MIT Press, 2000.

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

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

[8] A. D. Wilson and A. F. Bobick, "Paramteric hidden markov models for gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, 1999.

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

[10] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human-computer interaction," *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 415–422, 1997.

[11] A. Kobsa, J. Allgayer, C. Reddig, N. Reithinger, D. Schmauks, K. Harbusch, and W. Wahlster, "Combining deictic gestures and natural language for referent identification," in *Proceedings of the 11th International Conference on Computational Linguistics*, 1986.

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

[13] B. Fransen, V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski, "Using vision, acoustics, and natural language for disambiguation," in *Proceedings of the 2007 ACM Conference on Human-Robot Interaction (HRI '07)*, 2007.

[14] D. Perrett and N. Emery, "Understanding the intentions of others from visual signals: neurophysiological evidence," *Cahiers de Psychologie Cognitive*, vol. 13, no. 5, pp. 683–694, 1994.

[15] S. Langton, "The mutual influence of gaze and head orientation in the analysis of social attention direction," *The Quarterly Journal of Experimental Psychology: Section A*, vol. 53, no. 3, pp. 825–845, 2000.

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

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

[18] S. Baron-Cohen, "Precursors to a theory of mind: Understanding attention in others," in *Natural Theories of Mind*, A. Whiten, Ed. Oxford, UK: Blackwell Press, 1991, pp. 233–250.

[19] D. Baldwin and J. Moses, "Early understanding of referential intent and attentional focus: Evidence from language and emotion," in *Children's Early Understanding of Mind*, C. Lewis and P. Mitchell, Eds. New York: Lawrence Erlbaum Assoc., 1994, pp. 133–156.

[20] P. Bloom, "Mindreading, communication and the learning of names for things," *Mind and Language*, vol. 17, no. 1 and 2, pp. 37–54, 2002.