Proceedings of the 2003 IEEE/RSJ
Intl. Conference on Intelligent Robots and Systems
Las Vegas, Nevada · October 2003

Coupling Perception and Simulation: Steps Towards Conversational Robotics

Kai-yuh Hsiao Nikolaos Mavridis Deb Roy

Cognitive Machines Group
MIT Media Laboratory
20 Ames Street, Cambridge, MA 02142, USA
{eepness, nmav, dkroy}@media.mit.edu
http://www.media.mit.edu/cogmac

Abstract— **Human cognition makes extensive use of visualization and imagination. As a first step towards giving a robot similar abilities, we have built a robotic system that uses a perceptually-coupled physical simulator to produce an internal world model of the robot's environment. Real-time perceptual coupling ensures that the model is constantly kept in synchronization with the physical environment as the robot moves and obtains new sense data. This model allows the robot to be aware of objects no longer in its field of view (a form of "object permanence"), as well as to visualize its environment through the eyes of the user by enabling virtual shifts in point of view using synthetic vision operating within the simulator. This architecture provides a basis for our long term goals of developing conversational robots that can ground the meaning of spoken language in terms of sensorimotor representations.**

I. Introduction

Consider what would be required for a machine to understand the meaning of a sentence such as "Touch the heavy blue block that was on my left." Human understanding of the concepts underlying this statement draws upon a variety of cognitive abilities, including object permanence ("the block"), object properties and relations ("blue," "left"), category formation, theory of mind ("my"), working memory ("was"), visualization ("my left"), and knowledge of environmental affordances ("heavy"). Unless we endow robots with similar abilities, it is difficult to see how a robot can *truly* understand such a sentence, any more than a speech recognizer would "understand" its own transcriptions.

As a step towards this level of deep semantic understanding, we are developing a system with two tightly coupled components: a physical robot (called Ripley), and a physics simulator that serves as Ripley's "mental model" of the physical world.

We refer to the general problem of connecting the meaning of words and utterances to a robot's observations and actions as *language grounding*. Efforts have been made in the past to connect speech recognition systems to command-and-control robots (e.g., [3], [10], [13]; see [9] for a review of even more systems). Crangle and Suppes [4] present a detailed model of language mappings to robot perception and control that leads to a formal symbolic model of integrating grammar with semantics. This symbolic formal approach can be contrasted with work that emphasizes "subsymbolic" structures that link perception and action to the meaning of individual words. Along this latter approach, researchers have proposed detailed models for grounding the meanings of spatial terms [15], [16], color names [8], and verbs [1], [12], [21] in terms of sensorimotor associations. In our previous work, we adopted this approach of grounding words to develop several robotic and perceptually grounded systems that learn, understand, and generate natural spoken language [17], [18], [19].

There are important limitations to grounding words in terms of first-person sensorimotor associations. For example, the meaning of the word "left" in earlier works implicitly assumed a frame of reference from the robot's point of view. Even to use the word in simple conversation, however, a robot must be able to change points of view in order to see the difference between "my left" and "your left." This ability of a listener to assume a speaker's perspective is not limited to spatial perspective. For example, the speaker might hold different beliefs on the meaning of words (e.g., concrete words like "red"). The listener who is sensitive to such differences in word meanings and is able to accommodate them is more likely to communicate successfully. We think of these as forms of "modulation" of grounded meanings. The full spatial grounding of "left" is similar to the one from the first-person point of view, but the perspective shift operator serves to modulate the grounding so that it can be used more flexibly.

To explore the use of modulated grounded semantics, we have developed Ripley, a robotic manipulator with grasping capabilities and a multimodal sensory system, including stereo color vision, touch, and proprioception. Ripley's physical world is simple; it consists of a table with simple objects (such as beanbags and similar-sized

0-7803-7860-1/03/$17.00 © 2003 IEEE

items), and a human communication partner seated across the table. The robot's sensory system provides the required connections to reality, with cameras for visual object detection and sensors for joint positions, motor forces, and finger pressure.

As the robot moves about, sensory signals are used to drive a dynamics simulator of a 3-D world of rigid objects. Virtual objects, corresponding to the real, visually-detected objects, are instantiated on the basis of perceptual evidence. A virtual version of the robot follows the real robot through its motions and gestures, driven by data from the robot's joint position sensors. The position of the human user is registered, providing an external point of view to work with. The positions and properties of the objects are updated using new visual information, and when objects leave the field the simulator can continue to estimate their positions.

The motivation for connecting a physical system to a virtual simulator in real time comes chiefly from studies of mental imagery (see [7] for a review). These show not only that humans make extensive use of their visualization abilities for everyday tasks, but also that the visualization processes use the same sensory cortices that process real input. The simulation in our system provides an analogue to this cortical reuse, granting the ability to perceive real-world scenes in the same manner as imaginary scenes.

Apart from visualization, the simulator also provides a foundation for object permanence. Visually-detected objects are instantiated in simulation and persistently tracked. The simulator also includes a memory function that keeps a full history of world states and events and can be used to ground language that refers to the past, such as "the block that you were just holding." Finally, the simulator enables the system to view the environment from any spatial perspective. This last ability includes visualizing the world through the eyes of the human user. This ability to assume arbitrary spatial perspectives is critical for differentiating the meaning of phrases such as "my left" and "your left". More generally, it has been suggested that the ability to project the world through the perspective of another intentional agent plays a pivotal role in children's language acquisition [2] and thus is an important ability for conversational robots.

In summary, the simulator enables two major "perspective shifts": shifts in space (to view an object not in the physical line of sight, or to shift perspective), and shifts in time (to view a past scene). Combined with other perceptual processing, these abilities form a set of basis functions, representing elements of semantic concepts, with which words and linguistic phrases can be associated (see our previous work for examples of learning such associations). This in turn lays the groundwork for conversational robots.

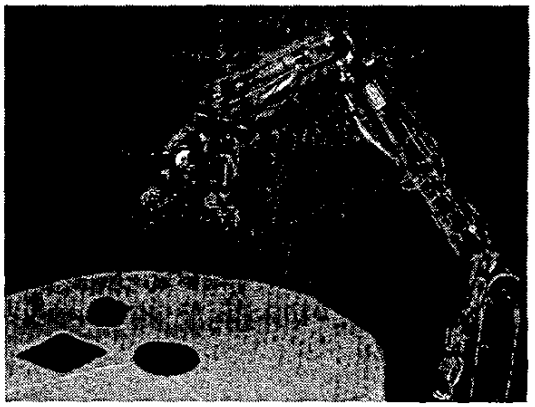

Fig. 1. Ripley. Joints at the base, elbow, and behind the head allow fluid movement about its tabletop domain. Notice also the cameras facing forward from the head, the gripper claw (in the open position), and the handle atop the head for direct user manipulation.

II. Ripley: An Interactive Robot

Ripley was designed specifically for the purposes of exploring questions of grounded language and interactive language acquisition. The robot has a range of motions that enables him to move objects around on a table-top placed in front of him, and to look around at the surrounding people and environment. In order to enable a meaningful sensorimotor grounding of verbs, Ripley's design included several specific elements:

- Ripley's body consists of a long arm with a gripper claw at the end for manipulating objects. This provides the potential for grounding verbs like "touch," and "lift."
- Ripley's "head" is also at the end of the arm and contains two cameras. This, along with its range of motion, causes the visual perspective to shift, making the notion of a shifting viewpoint an integral part of the system.
- Ripley has compliant joints and training handles, which enable the human user to demonstrate gestures while narrating to provide linguistic associations.

A. Structure and Actuation

The robot (see Fig. 1) has seven degrees of freedom (DOF's). Each DOF other than the gripper is actuated by series-elastic actuators [14], which enable the force applied by the motors to be controlled directly, in contrast to motors which are controlled by speed. The use of series-elastic actuators gives Ripley the ability to precisely sense the amount of force that is being applied at each DOF, and leads to compliant motions, in which the robot is aware of external forces and can choose how strongly to compensate.

B. Basic Motion Control

Motion control in Ripley is inspired by studies of motor force fields in frogs [11]. In essence, frogs' limbs are controlled by internally-represented force fields, which cause a given limb to converge on a single point, and this convergence point is moved smoothly along the planned trajectory of the limb. As a rough approximation to this method, a position-derivative control loop is used to track a target point that transits smoothly from the starting point of a motion gesture to the end point. Forces are computed every five milliseconds, based on trajectories computed by the host computer.

Manual training of the robot allows the user to demonstrate new gestures by moving Ripley like a puppet. This requires that the robot be easy to move. To this end, we set up a training mode in which Ripley exerts just enough force at each joint to counteract gravity, using a forward kinematics model to predict the force of gravity on each joint, which is then cancelled by forces from the actuators. Motion interpolation algorithms are used to generalize from trained trajectories to novel motions as dictated by new perceptual contexts.

C. Sensory System and Visual Processing

Complementing Ripley's motor system is a perceptual system, with two color video cameras, a two-axis tilt accelerometer (for sensing gravity), and two microphones mounted in the head. Force sensitive resistors provide a sense of touch on the inside and outside surfaces of the gripper fingers.

One of the most important sets of sensors is embedded in the actuators. As described above, the actuators are force-controlled, meaning that the amount of force being applied at each joint can be sensed by virtue of being controlled. Additionally, each DOF is equipped with absolute position sensors, providing data for all levels of motion control and for maintaining the anti-gravity mode.

The vision system is responsible for detecting objects in the robot's field of view. Background and foreground Gaussian color models are applied to detect connected regions at a frame rate of 10 Hz. The detected visual regions are passed along to the "Objecter" module, described below, which integrates region analyses over time to determine the presence and properties of objects in the scene. This vision model is very simplistic and requires not only that objects be single-colored, but also that they be in the plane of the table so the simulator can adequately infer their 3-D positions (more detail below). However, this system is sufficient for examining our mental model (which is the primary purpose of this paper) and we are starting to look into more sophisticated 3-D vision algorithms with more complex object models.

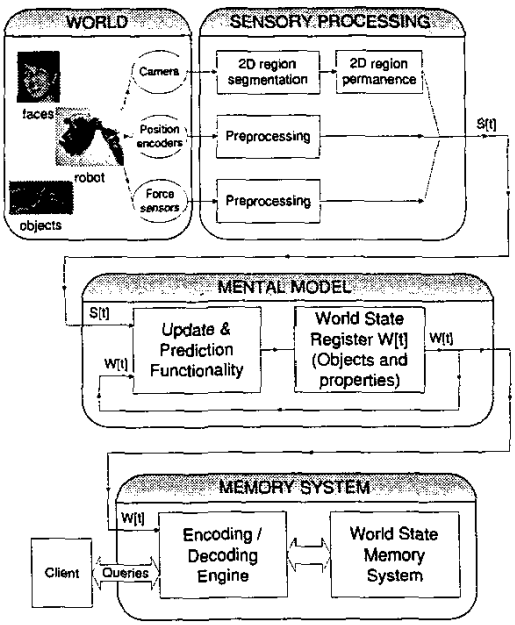

Fig. 2. Block diagram of our system. Sensory information is processed and passed to the simulator's mental model and memory.

III. A Perceptually-Driven "Mental Model"

At this point, we are ready to feed the sensory information into a "mental model." In our approach (see Fig. 2), Ripley's simulator (SimRip) integrates real-time information from its visual and proprioceptive systems to construct an internalized mental model, which tries to best explain the history of sensory data that has been observed.

Brian Cantwell Smith has argued for the fundamental importance of the ability to internalize and track objects, in order to retain awareness of them even when direct perception fails. In his words, "the retraction of responsibility into the [subject] to compensate for the loss in effective coupling [with the object] - this is the origin of reasoning, representation and syntax" [22].

Aside from object permanence, the simulator also allows Ripley to envision its world through the eyes of the human user. After the persistent objects have been instantiated, it is a simple matter to have the point of view rotated across the table to the perceived location of the user. Extraction of object features based on a dynamic viewpoint will allow the learning of novel aspects of conversation, as discussed previously.

A. Physical Simulation

In the heart of SimRip lies the ODE rigid body dynamics simulator [23]. ODE handles masses of arbitrary geometries and updates the world in discrete time steps based on Newtonian mechanics. Support for collisions,

joints (restrictions of relative motion), soft second order constraints, and friction is provided by the ODE engine. Passive objects and human faces are modeled as single spherical ODE objects, while vRip's body (Ripley's self image living in SimRip) is modeled as a configuration of seven cylindrical links plus a rectangular head, with dimensions and masses approximating reality. The properties of these objects (passive, face, and vRip's body) form the state of the mental model.

B. Coupling Perception to the Mental Model

Before using the visual input to inform the internal model, we must correct for noise in the vision system. Even with the simplistic environment, the object detector is still confused by camera noise, shadows, head motion, and model simplicity. All the visual attributes are subject to some noise, and a persistently detected object may even disappear for a frame or two. Furthermore, the motion of Ripley's head and of the objects causes significant changes in the size, color, and position of objects.

To compensate for such motion and noise, the object detector passes its input to the "Objecter" module, which filters out noise by tracking objects from frame to frame, within the 2-D field of view. To do this, it keeps a running database of objects that have been encountered.

Visual objects consist of a size, a position, and a color. By using a distance metric on these features to compare objects in the database to objects passed from the vision system, it finds a minimal-distance mapping between the database and the visual frame. Any object that is clearly new (i.e., is very distant from objects in the previous frame) is labeled as such and added to the database. Objects in the database which have not been seen in a set number of frames are deleted. Finally, when an object in the database has been seen consistently enough, it is added to the list of objects to be instantiated in the simulator.

By deleting and instantiating objects only after several frames, the Objecter thus serves to reduce noise and provide a sort of hysteresis for object creation, helping to offset the effects of brief occlusion. It also allows the visual input to track an object persistently between frames, despite noise and motion. On the other hand, this tracking is done only in a 2-D domain, and objects that leave the field of view are soon forgotten. For this reason, the Objecter is similar to the human analogues of sensory memory and visual tracking. Higher-level, 3-D object permanence is reserved for the simulator itself.

The other aspect of perception coupled to the simulation is the output of the proprioception system. This consists of a vector of seven angles uniquely determining the physical configuration of the robot, and a vector of forces applied by the actuators (useful for a sensation of "difficulty" of movement, e.g. in weighing or measuring softness of objects).

These data streams are sent via network to the simulator program. The proprioceptive information is used to update the position of the virtual robot, and the processed visual information is used to update the positions and properties of the objects maintained within the simulation.

C. Dynamics and Memory in the Mental Model

Prediction for the physical part of objects is already provided by ODE, using numerical integration with Newton's laws. Thus, thanks to ODE, an object that was instantiated, but which has left the visual field, will continue moving if it was last seen moving, and it may even move with a non-constant velocity if friction or other forces are taken into account. For instance, if an object's trajectory passes through another object that was last seen stationary and is also out of view, then the collision will be predicted.

Another problem indirectly simplified by the simulator is that of projecting the 2-D object data from the vision system into 3-D space. Because the positions of the robot's head and the table are modeled accurately in the simulator, it is possible to estimate a 3-D position for objects based only on a 2-D input, which is ideal for our current, very simplistic vision model. When we transition to a more sophisticated 3-D vision system, this estimation will no longer be necessary.

Using the simulator does not simplify everything, though. Using the sensory data to actually update the simulation model requires additional processing, and several assumptions. A static viewpoint with unmoving objects is easy to model, but because objects can enter and leave the field of view, due either to motion of the objects or motion of the robot's head, a more detailed model of permanence is needed.

We begin with the assumption that there are very few "magical disappearances" [24]. Moving objects are expected to enter and exit through the borders of the field of view, and upon leaving they should be assigned the appropriate velocity and simulated in the absence of sensory input. Thus, even though the visual system (via the Objecter) stops tracking the object, the simulator continues to be aware of its existence. A few problems arise with this, such as objects apparently shrinking as they leave the field of view, because the part of the object that is visible gradually decreases. Also, objects moving quickly relative to the 10 Hz frame rate seem to magically disappear. Simple heuristics partially compensating for these have been implemented.

Moving viewpoints present further complications. Our current implementation simply ignores object input while the robot is moving, but we are working on a system to stabilize the simulator world dynamically during robot motion. For instance, when an object enters the field of view during a perspective shift, it could either be a new object or an object previously tracked by the simulator.

To help determine this, a module similar to the Objecter, but working in 3-D, has been written in the simulator to perform the necessary comparisons.

In our current implementation of the simulator, we made several other simplifying assumptions as well. We already mentioned the use of the simulator to use 2-D data from only one camera to instantiate 3-D objects based on the position of the head and the table. Likewise, the 3-D position of the user's head can also be estimated by projecting 2-D position information from a stock face detector onto a surrounding sphere.

Also, objects are assumed to be spherical, and color is assumed to be homogeneous across an object. See Fig. 4 for an example of object instantiation, corresponding to the visual input shown in Fig. 3. Furthermore, Fig. 6 and Figure 7 show two timesteps of an object being dragged across the table, as in Fig. 5.

Finally, we have a simple memory system, which stores all the past states of the mental model. Thus, the history of objects can be retrieved, examined, and replayed through different viewpoints, providing the foundation for grounding such constructs as verb tenses.

IV. Towards Grounding Conversational Language

Let us return to the spoken utterance we considered earlier, "Touch the heavy blue block that was on my left." Given the coupled robot-simulator architecture that has been described, we can now sketch how the words in this utterance might be grounded in terms of sensorimotor grounded structures in this system. "Touch" is grounded in a procedure that reaches towards targeted objects by interpolating human-trained motion trajectories. "Heavy" specifies a range of values from a weighing procedure. To weigh an object, Ripley grasps and moves the object up and down, gauging the force applied to determine the relative weight. In contrast to visual properties such as size and color, finding the weight of an object inherently requires motor interactions (similar to Gibson's notion of affordances [5]). "Blue" is grounded in terms of the color space from Ripley's visual system. "Was" triggers an index back in time, which is supported by the event-based memory of Ripley's simulator. Finally, "my left" can be grounded through a perspective shift to the human communication partner's physical point of view.

V. Ongoing Work

We are currently in the process of integrating speech processing [20], spatial language processing [6], and associative learning [19] subsystems into the architecture described here. In addition, we plan to pursue several other directions. These include probabilistic representations of partial knowledge of object properties in the simulator, enhancement of the vision system to deal with partial occlusion, more detailed modeling of object geometry, and grounding of manipulation verbs.

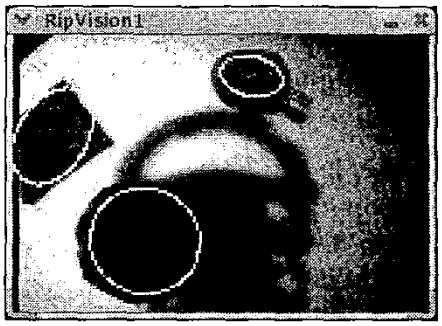

Fig. 3. A physical scene as seen through Ripley's camera.

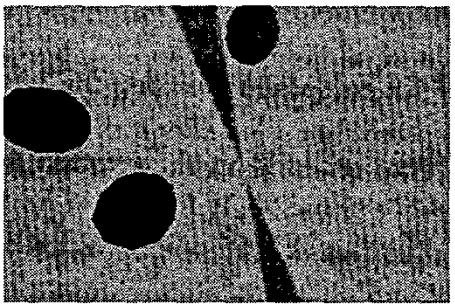

Fig. 4. The same scene, recreated in the simulator model. The triangles represent the axis of the table.

Using our coupled simulation system, we have developed a robot that can maintain simple object permanence and imagine its world through the eyes of its user. This provides it with a structured foundation upon which it can ground common phrases such as "on my left." These components, along with the others now being developed, will hopefully provide an enriched basis for robots that can engage in fluid, situated conversations with people.

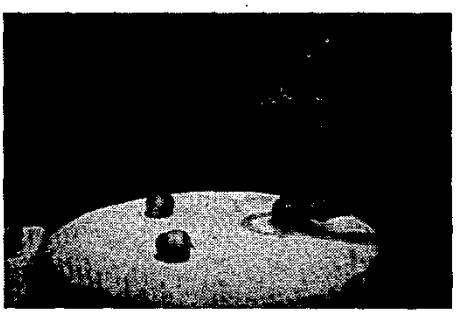

Fig. 5. The user pulls a ball across the table.

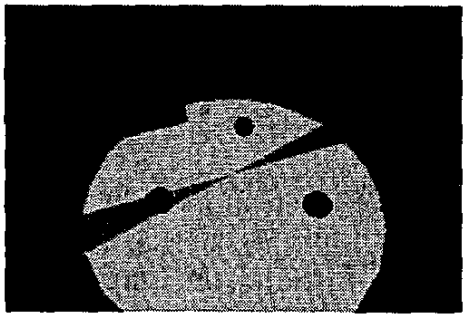

Fig. 6. The ball (on the left) in its initial simulated position.

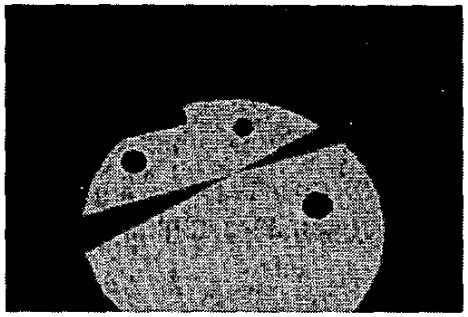

Fig. 7. The ball, pulled towards the table's edge.

ACKNOWLEDGMENTS

Thanks to Ben Krupp and Chris Morse for their help constructing Ripley, and to Niloy Mukherjee for implementation of the vision system.

VI. REFERENCES

[1] D. Bailey. *When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs*. PhD thesis, Computer science division, EECS Department, University of California at Berkeley, 1997.

[2] P. Bloom. Mindreading, communication and the learning of names for things. *Mind & Language*, 17:37–54, 2002.

[3] M. K. Brown, B. M. Buntschuh, and J. G. Wilpon. Sam: A perceptive spoken language understanding robot. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(6), 1992.

[4] C. Crangle and P. Suppes. *Language and Learning for Robots*. CSLI Publications, Stanford, CA, 1994.

[5] J. J. Gibson. *The Ecological Approach to Visual Perception*. Erlbaum, 1979.

[6] P. Gorniak and D. Roy. Grounded semantic composition for visual scenes, forthcoming, 2003.

[7] S. M. Kosslyn, G. Ganis, and W. L. Thompson. Neural foundations of imagery. *Nature Reviews: Neuroscience*, 2:635–642, 2001.

[8] J. M. Lammens. *A computational model of color perception and color naming*. PhD thesis, State University of New York, 1994.

[9] L. S. Lopes and J. H. Connell. Semisentient robots: Routes to integrated intelligence. *IEEE Intelligent Systems*, 16:10–14, 2001.

[10] P. McGuire, J. Fritsch, J. Steil, F. Roethling, G. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter. Multimodal human-machine communication for instructing robot grasping tasks. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2002.

[11] F. A. Mussa-Ivaldi and E. Bizzi. Motor learning through the combination of primitives. *Philosophical Transactions of the Royal Society of London*, 355:1755–1769, 2000.

[12] S. Narayanan. *KARMA: Knowledge-based active representations for metaphor and aspect*. PhD thesis, University of California Berkeley, 1997.

[13] D. Perzanowski, A. Schultz, W. Adams, K. Wauchope, E. Marsh, and M. Bugajska. Interbot: A multi-modal interface to mobile robots. In *Proceedings of Language Technologies 2001*, Carnegie Mellon University, 2001.

[14] J. Pratt, B. Krupp, and C. Morse. Series elastic actuators for high fidelity force control. *Industrial Robot*, 29(3):234–241, 2002.

[15] T. Regier. *The human semantic potential*. MIT Press, Cambridge, MA, 1996.

[16] T. Regier and L. Carlson. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology*, 130(2):273–298, 2001.

[17] D. Roy. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3), 2002.

[18] D. Roy. Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, 2003 (forthcoming).

[19] D. Roy, P. Gorniak, N. Mukherjee, and J. Juster. A trainable spoken language understanding system for visual object selection. In *International conference of spoken language processing*, 2002.

[20] D. Roy and N. Mukherjee. Semantic priming in speech understanding using visual context, forthcoming, 2003.

[21] J. Siskind. Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. *Journal of Artificial Intelligence Research*, 15:31–90, 2001.

[22] B. C. Smith. *On the Origin of Objects*. Bradford Books, 1996.

[23] R. Smith. Ode: Open dynamics engine. http://q12.org/ode/.

[24] K. Wynn and W.-C. Chiang. Limits to infants knowledge of objects: The case of magical appearance. *Psychological Science*, 9:448–455, 1998.