Mental Imagery for a Conversational Robot

Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis

***Abstract*—To build robots that engage in fluid face-to-face spoken conversations with people, robots must have ways to connect what they say to what they see. A critical aspect of how language connects to vision is that language encodes points of view. The meaning of *my left* and *your left* differs due to an implied shift of visual perspective. The connection of language to vision also relies on object permanence. We can talk about things that are not in view. For a robot to participate in situated spoken dialog, it must have the capacity to imagine shifts of perspective, and it must maintain object permanence. We present a set of representations and procedures that enable a robotic manipulator to maintain a "mental model" of its physical environment by coupling active vision to physical simulation. Within this model, "imagined" views can be generated from arbitrary perspectives, providing the basis for situated language comprehension and production. An initial application of mental imagery for spatial language understanding for an interactive robot is described.**

***Index Terms*—Active vision, grounding, language, mental imagery, mental models, mental simulation, robots.**

I. Situated Language Use

In using language to convey meaning to listeners, speakers leverage situational context [1], [2]. Context may include many levels of knowledge ranging from the details of shared physical environments to cultural norms. As the degree of shared context decreases between communication partners, the efficiency of language also decreases since the speaker is forced to explicate increasing quantities of information that could otherwise be left unsaid. A sufficient lack of common ground can lead to communication failures.

If machines are to engage in meaningful, fluent, situated spoken dialog, they must be aware of their situational context. As a starting point, we focus our attention on physical context. A machine that is aware of where it is, what it is doing, the presence and activities of other objects and people in its vicinity, and salient aspects of recent history, can use these contextual factors to interpret natural language.

In numerous applications of spoken language technologies such as talking car navigation systems and speech-based control of portable devices, we envision machines that connect word meanings to the machine's immediate environments. For example, if a car navigation system could see landmarks in its vicinity based on computer vision, and anchor descriptive language to this visual perception, then the system would have a basis for generating contextually appropriate directions such as "Take a left turn immediately after the large red building." Consider also an assistive service robot that can lend a helping hand based on spoken requests from a human user. For the robot to properly interpret requests such as "Hand me the red cup and put it to the right of my plate," the robot must connect the meaning of verbs, nouns, adjectives, and spatial language to the robot's perceptual and action systems in a situationally appropriate way.

Our current work is part of a larger effort to develop a conversational interface for an interactive robot (see also [3]–[6]). The development of such a robot is of practical interest in domains ranging from space exploration (e.g., [7]) to assistive aids (e.g., [8]). Furthermore, we believe that lessons learned from developing robotic interfaces may have impact in numerous other natural language processing domains.

A necessary step toward creating situated speech processing systems is to develop representations and procedures that enable machines to ground the meaning of words in their physical environments. In contrast to dictionary definitions that represent words in terms of other words (leading, inevitably, to circular definitions), grounded definitions anchor word meanings in nonlinguistic primitives. Assuming that a machine has access to its environment through appropriate sensory channels, language grounding enables machines to link linguistic meanings to elements of the machine's physical world.

Interest has grown in the computational representation and acquisition of word meaning grounded in vision [9]–[18] and motor action [19]–[21]. This line of research, in addition to making contributions to theoretical aspects of lexical semantics and cognitive modeling, has practical relevance for building situated human-machine communication systems. A limitation of this previous work, however, is the assumption of a fixed, first-person visual frame of reference.

Our approach departs from the assumption of camera-grounded fixed perspective by introducing an implemented model of *mental imagery* driven by active vision. Mental imagery enables grounding of spatial language that cannot be handled under fixed-perspective assumptions. To understand the difference between *behind me* and *behind you*, the listener must factor points of view into the language comprehension process. Speakers must similarly take into account listeners' points of view to produce clear, unambiguous language. Simpler solutions such as in-plane rotation of images to correct for perspective will not work in general, since full three-dimensional (3-D) changes of perspective are required in many situations. Furthermore, mental imagery enables anticipation of visual occlusions which are view dependent and cannot be predicted through image rotations. Our approach also introduces object permanence so that language can bind to objects that are not in direct view of the system's camera. As a result, the system can understand and generate language about objects which are not physically in the camera's sight.

Manuscript received April 14, 2003; revised September 3, 2003 and December 12, 2003. This paper was recommended by Associate Editor H. Zhang.

The authors are with the Massachussetts Institute of Technology (MIT) Media Laboratory, MIT, Cambridge, MA 02142 USA (e-mail: dkroy@media.mit.edu).

Digital Object Identifier 10.1109/TSMCB.2004.823327

We first introduce our notion of mental imagery and its role in language use. We then present details of an implementation of a computer vision driven mental model that is used to generate mental imagery. We conclude by presenting an application of language understanding grounded in mental imagery. Although we build on earlier work on visually-grounded language cited above, this work makes a significant departure by defining a new way to connect language and vision that is better able to address the needs of situated language processing.

II. Mental Models and Mental Imagery: Where Language and Vision Meet

A key aspect of human perception is that it is active. We cannot move without affecting our senses, and in order to perceive, we must coordinate our movements. In the realm of visual perception, movement of the head and body leads to apparent motion in the visual field, and the appearance/disappearance of objects from the field of view. Yet, we are able to conceptualize the world as stable, maintain object permanence in the face of appearances and disappearances, and differentiate self-motion from motion in the environment.

We adopt the term *mental model* to refer to the conceptual structures that represent a stabilized version of reality, essentially a "cache" of the external world as projected through the observer's perceptual system. The idea of mental models is well established in the cognitive science literature (cf. [22]) although it is more typically used to describe offline cognitive processes where perception is not directly driving the construction and updates of the mental model. In our approach, perceptually driven mental models provide a level of abstraction above low-level vision that is appropriate for connecting to language (along these lines, see also [23]).

We also adopt the term *mental imagery* to refer to images that are generated by imagining viewpoints within a mental model. The Stanford Encyclopedia of philosophy [24] defines mental imagery as: *Experience that resembles perceptual experience, but which occurs in the absence of the appropriate stimuli for the relevant perception [25], [26]. Very often these experiences are understood by their subjects as echoes or reconstructions of actual perceptual experiences from their past; at other times they may seem to anticipate possible, often desired or feared, future experiences.*

Our use of the term extends this definition since we are concerned with representations and processes that are active during actual perceptual experience. We choose to use the same term for both cases based on our intuition that many of the same processes used for online perception are also used for offline reconstruction and reasoning (see [27], [28] for psychological arguments in support of this view).

Language refers to the stabilized conceptualization of the world provided by mental models and imagery – we do not talk of objects as being in motion when we know that the apparent motion was caused by our own movements. We also talk about objects that are out of view if we are certain of their location. Moreover, spatial language in situated dialogs assumes a point of view that will depend on how the speaker decides to express herself. Perspective taking has long been studied in psychology, leading to a large literature on the subject including the developmental studies of Piaget [29]. Tversky provides a useful taxonomy of spatial thinking [30]. In her analysis, basic kinds of frames of reference that humans use to conceptualize space include space of the body (body parts), space around the body, and the space of navigation. Here, we primarily address computational representations of space around the body of a robot. The ability to shift perspectives is also related to aspects of space in navigation, although verbal interaction with a mobile robot (e.g., [31]) addresses the latter more directly.

Using Miller and Johnson–Laird's terminology [32], speakers may assume a first-person *deictic* frame of reference (e.g., "on my left"), or alternatively an *intrinsic* perspective (e.g., "on your left" or "in front of the house"). Intrinsic expressions occur when spatial terms are used to indicate positions relative to entities that have intrinsic parts (e.g., houses have fronts and backs) and may thus serve as the basis for spatial frames of reference. One way for a listener to interpret the meaning of deictic references, and the approach that we have explored in our computational model, is to use mental imagery to visualize the shared scene from the speakers point of view, and within this shifted frame, interpret spatial expressions. In other words, the phrase "on my left" is decomposed into two parts, "my," and "on – left". The "my" part triggers a shift of perspective to the speaker's point of view. Similar strategies can be used within this framework to shift perspectives according to intrinsic frames of reference.

Imagining how a shared environment looks from another's perspective is often crucial to effective communication. If an object is in view to speaker S, but not listener L, S should take this factor into account when referring to the object. If S knows that L cannot see an apple because it is behind a basket, S might say "the apple behind the basket" rather than just "the apple". If the apple is in view to both parties, the former description would seem odd since it specifies unnecessarily redundant details.

To summarize, language cannot be grounded directly in first-person visual representations. Language must instead be grounded through some other representational layer which provides a stable view of the environment in spite of self-motion. This middle ground also enables speakers and listeners to imagine each other's point of view, a necessary precondition for natural situated spoken dialog.

With this motivation in mind, we present an architecture for actively constructing mental models.

III. Physical Embodiment: Ripley

Our current experiments are based on a robotic manipulator named Ripley (Fig. 1). Ripley has seven degrees of freedom (DOFs), enabling it to manipulate objects in a three-foot radius workspace. The robot may be thought of as an articulated torso terminating with a head that includes its "mouth" (a one DOF gripper).

Ripley has been designed to explore situated, embodied spoken language use. In contrast to our previous robots [13], [33], Ripley is able to use its gripper to manipulate small objects, paving the way for grounding verbs related to manipulation actions. The robot's range of motions enable it to

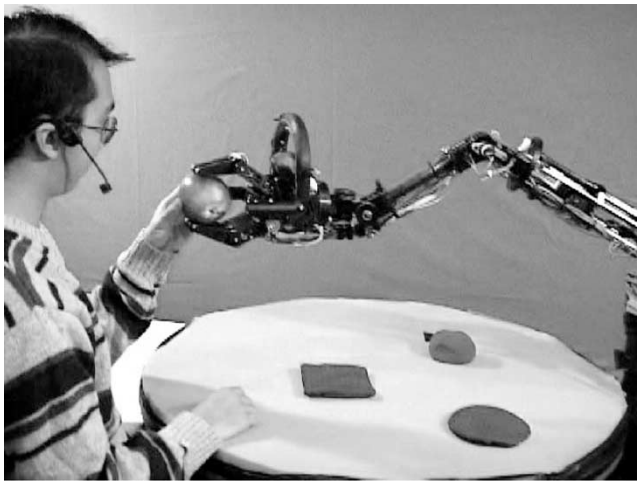

Fig. 1. Ripley hands an apple to its human communication partner in response to the phrase "Hand me the thing on your left."

examine objects through vision and touch. Ripley is also able to look up and make "eye contact" with its human communication partner. This behavior plays a functional role since Ripley must keep track of the position of the partner in order to understand relative spatial reference. Eye contact is also, of course, important for engaging in natural face-to-face dialog.

Most of Ripley's sensors are in its head, including two color video cameras, two microphones, touch sensors, and an inertial sensor for gravity. Additional proprioceptive (position and force) sensors are placed on each joint. In our current work, only one of the cameras is used for visual input.[1] The placement of the camera on the mouth simplifies grasping since visual servoing can be used to guide the gripper to objects. However, the placement also leads to constant changes in the robot's field of view since any motion of the torso affects the camera. For this reason, Ripley provides an excellent platform for developing mechanisms for mental imagery.

Low-level motor control is achieved by computing trajectories of target joint configurations. An elastic force model loosely inspired by motor force fields in biological motor control [34] is used to provide compliant motion control [20]. Higher level motor control directives are issued from a planning mechanism that is driven by task specific criteria.

Low-level visual processing relies on color-based separation of objects from a known (fixed) background (the image processing methods are described in [33]). The vision system generates a set of foreground regions at a rate of 15 Hz. These region sets are passed to an object permanence module which integrates region sets over time to determine the presence and properties of objects in the scene. As we describe in the next section, the object permanence module uses the robot's joint configurations to compensate for view points in order to maintain a view-independent model of object locations. A face detector [35] searches for faces in the visual field. Faces are treated specially, leading to a model of the communication partner's location in the robot's mental model.

[1]In ongoing work, we are introducing depth perception based on stereo visual input.

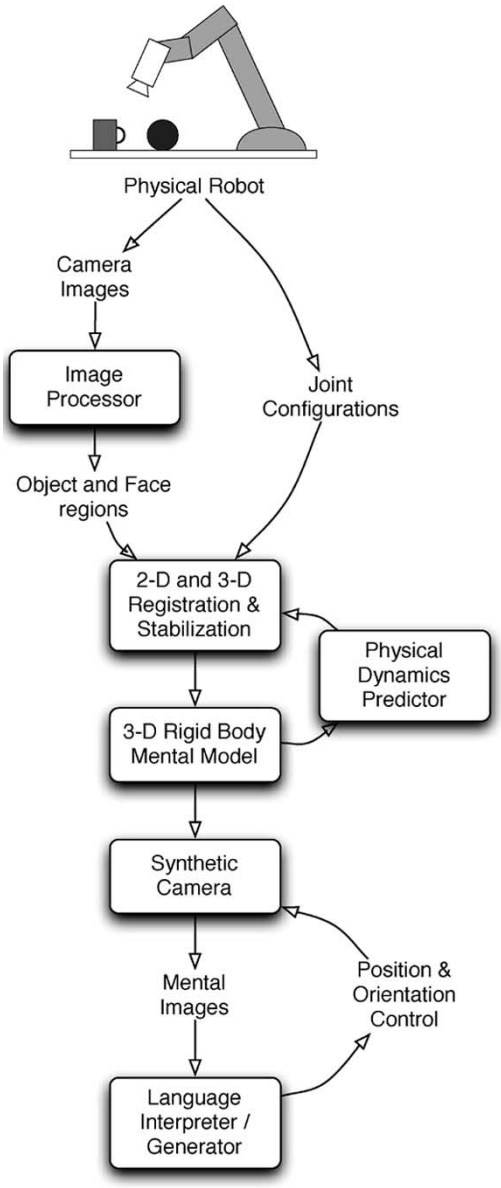

Fig. 2. Architectural overview: Active vision drives the construction and maintenance of a mental model. Synthetic mental images from the mental model are linked to language.

IV. Mental Model and Mental Imagery

As we move around our direction of gaze, objects come in and out of sight, but our conception of objects remains stable. Fig. 2 provides an overview of Ripley's mental model and imagery architecture that registers and stabilizes sensory data from the robot's moving camera. We begin with an overview of the architecture. Subsequent sections highlight technical details of the implementation.

Ripley's camera provides a constant stream of images to the image processor. The image processor finds foreground regions (typically corresponding to the location of objects on the robot's work surface) and face locations which are relayed, at the 15-Hz frame rate, to a registration and stabilization module. This module constructs and maintains a 3-D model of objects in the environment. The robot's joint configuration is used to perform projective transforms on incoming images so that a

3-D model can be created out of multiple two-dimensional (2-D) views.

The mental model is a 3-D model consisting of a set of rigid body objects, and represents Ripley's belief of the state of the world. The registration and stabilization module acts as a sophisticated hysteresis function to smooth sensory data. Persistent perceptual evidence for the presence, movement, or disappearance of objects drives updates in the mental model.

A physical dynamics estimator is used to provide predictions of where objects should be in incoming image frames, given the current state of the mental model and knowledge of Newtonian physics. This predicted model is used to align perspective-dependent camera image regions with the contents of the 3-D mental model. The contents of the mental model can be used to generate synthetic images using a synthetic camera and standard projective computer graphics techniques.

The language processor receives these synthetic images as a basis for grounding semantics. The language processor has control over the position and orientation of the synthetic camera. To interpret spatial language, the synthetic camera can be positioned to simulate either the robot's or the human partner's point of view.

A. Representation of Mental Model State

The physical environment is modeled by a set of rigid 3-D objects which includes: 1) a model of Ripley's own body; 2) a built in model of the workspace table; 3) a physical model of the human communication partner's body; and 4) objects situated on the work surface. The complete state of the mental model is captured at any moment by the descriptions of all objects. Each object in the model is fully described by its position, orientation, shape, color, mass, and velocity. The self-model consists of a set of four cylindrical blocks connected by swivel joints to approximate the shape and range of positions of the physical robot. The physical model of the human partner is currently a simple sphere which is used to position synthetic cameras to obtain the human's point of view.

B. Dynamic Prediction

The open dynamics engine (ODE) rigid body dynamics simulator [36] is used to predict future states in the mental model. ODE is an open source Newtonian physics simulator that operates on 3-D rigid body models. At a rate of 10 Hz, the state of the mental model is copied into ODE, ODE is executed to generate a prediction for the next time step, and this predicted state is integrated with perceptual evidence from Ripley's camera and joint sensors to update the state of the mental model (see below). ODE thus provides two main functions within our system: collision detection and dynamics simulation.

C. Ripley's Physical Self-Model

The model of the robot's body is controlled by a simulated position-derivative motor controller similar to the controller used in the physical robot. At each update cycle in the model, joint angles of the virtual robot are compared to the angles of the physical robot. For each joint, if the difference in angles is greater than a preset threshold, then an appropriate force is applied to align the corresponding virtual joint. In effect, the virtual joint tracks the associated DOF of the physical robot. Since only angle differences above a threshold lead to virtual forces, low-level jitter in the physical robot is attenuated in the mental model.

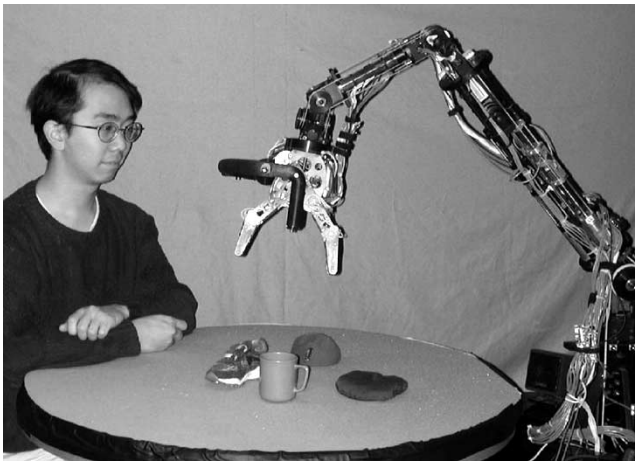

Fig. 3. Ripley looks down at the tabletop with four objects in view.

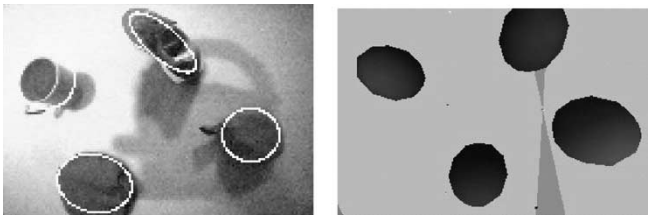

Fig. 4. Visual regions and corresponding simulated objects in Ripley's mental model corresponding to the view from Fig. 3. The white ellipses in the left image indicate the output of the region analysis routines of the vision system. The objects in the simulator on the right are actually spherical, but appear elliptical due to optical warp of the synthetic viewpoint generated by the simulator.

D. Coupling Active Vision to the Mental Model

A primary motivation for developing the mental model is to register, stabilize, and track visually observed objects in Ripley's environment. To address these needs, an object permanence module, called the *Objecter*, bridges the incoming stream of input from the image analysis module to the contents of the mental model (for other approaches to perceptually coupled simulation, see [37]–[39]). When an image region is found to stably exist for a sustained period of time, an object is instantiated by the Objecter in the mental model. The color and position of the object are determined from the visual input. It is only at this point that Ripley becomes "aware" of the object and is able to talk about it. If Ripley looks away from an object so that the object moves out of view, a representation of the object persists in the mental model. When a physical object is removed from Ripley's workspace, persistent perceptual evidence of its disappearance causes the object to be deleted from the model.

Fig. 3 shows an example of Ripley looking over its workspace with four objects in view. In Fig. 4, the left image shows the output from Ripley's head-mounted camera, and the right image shows corresponding simulated objects that have been registered and which are being tracked.

The Objecter consists of three components: a 2-D-Objecter, 3-D-Objecter, and 2-D-to-3-D resolver. The 2-D-Objecter tracks 2-D visual regions generated by the vision system. The 2-D-Objecter implements a hysteresis function which detects 2-D visual regions that persist over time, and resolves intra-frame region correspondences, assigning unique IDs to persistent regions. The 2-D-to-3-D resolver module follows, which calculates the position and pose of prospective 3-D objects, based on the persistent 2-D regions it is fed. Due to a lack of depth information, the resolver relies on projective geometry and the assumption that objects are in contact with Ripley's table. The 3-D-Objecter brings the prospective 3-D objects into correspondence with those already existing in the mental model, and decides whether and how to create, update or delete objects in the mental model.

As Ripley moves (and thus changes its vantage point), the 2-D-Objecter continues to track visual regions until they leave the field of view. However, updates to the 3-D mental model are not performed while Ripley is in motion. This simplifies the process of tracking objects and leads to greater model accuracy. Overall, as a coupled pair, the 2-D and 3-D Objecter maintain correspondence of objects across time, enabling tracking and object persistence in spite of perceptual gaps, noise, and spatial reorderings of the objects.

More precisely, the output of the image processing module at each time step is a set of N visual regions, $V[t] = \{R_1^v[t], R_2^v[t], \ldots, R_N^v[t]\}$. In general, the ordering of regions within V is arbitrary since the vision system finds regions in each frame of video independent of knowledge of previous frames. Thus, there is no guarantee that $R_i^v[t]$ will correspond to $R_i^v[t+1]$.

To obtain correspondence of regions over time, the 2-D-Objecter maintains its own set of regions which are candidates for being output to the 3-D-Objecter. We denote the candidate region set as $O[t] = \{R_1^o[t], R_2^o[t], \ldots, R_M^o[t]\}$. The purpose of the 2-D-Objecter is to maintain correspondence between $R_i^o[t]$ and $R_i^o[t+1]$. To maintain region correspondence, we define a tunable distance metric between two visual regions as

$$d(R_i, R_j) = \alpha d_p(R_i, R_j) + \beta d_s(R_i, R_j) + (1 - \alpha - \beta) d_c(R_i, R_j) \quad (1)$$

where $d_p()$ is the Euclidean distance between the centroids of the regions, $d_s()$ is the difference in size (number of pixels) of regions, and $d_c()$ is the difference in average RGB color of the regions. The tuning parameters α and β are scalar values such that $(\alpha + \beta) \leq 1$. They are used to set the relative emphasis of the position, size, and color properties in comparing regions.

When Ripley moves to a new vantage point, the 2-D-Objecter candidates are initialized by copying the output of the vision system ($O \leftarrow V$) so that a candidate is created corresponding to each region in the current visual analysis frame. A confidence value, $R_i^o[t]$.conf, is assigned to each candidate and initialized to 0. At each successive time step, a new region set is generated by the vision system. The 2-D-Objecter attempts to put each region in V into one-to-one correspondence with each candidate in O such that the total distance using (1) between paired regions is minimized. In general, the number of visual regions N and 2-D-Objecter candidate regions M will not be equal. The alignment process aligns the $\min(N, M)$ subset of regions. After the optimal alignment is found, only those whose distances resulting from the match are below a maximum allowable distance threshold are accepted. The confidences of candidate regions that are aligned to regions from V are updated (increased) using a rule similar to an infinite impulse response filter, assuming positive input of unit magnitude. Thus, confidence values never reach an upper bound of 1.0. If $N > M$, at most $(N - M)$ new candidates are instantiated in the 2-D-Objecter, each with confidence set to 0. If $N < M$, then the confidence of, at minimum, $(M - N)$ unaligned candidate regions is updated (decreased) by a similar rule, driven by a negative input of unit magnitude. At the end of this alignment and confidence update process, the properties of the matched or newly instantiated regions from the 2-D V are copied into O. The unmatched candidate regions retain their previous properties, and any of them for which $R_i^o[t]$.conf < 0 are destroyed.

The output of the 2-D-Objecter at each time step is the subset of candidate regions for which the confidence level is greater than Conf_{MIN}. In the current implementation $\text{Conf}_{\text{MIN}} = 0.9$. Each newly instantiated candidate region is assigned a unique ID. These IDs are persistent over time, thus implementing region tracking. Smoothly moving objects are tracked by the 2-D-Objecter. When an object is removed from a scene, the confidence value of the corresponding candidate region will start dropping from the maximum value of Conf_{MAX}. As soon as the confidence drops below Conf_{MIN}, it stops being output. This use of confidence values and thresholds implements a hysteresis function that requires persistent visual evidence before either instantiating or destroying regions.

The 3-D-Objecter uses projective geometry to infer the position of objects in 3-D space based on 2-D regions. Given the position and orientation of Ripley's camera, the 2-D regions are linearly projected in 3-D until the projection lines intersect Ripley's work surface. The location of the surface, a round tabletop, is built into the initial state of the mental model. Thus, Ripley's perceptual input is not necessary for establishing the presence of the table.

Interaction between the 2-D- and 3-D-Objecter proceeds as follows. Each time Ripley moves, the 3-D-Objecter ignores output from the 2-D-Objecter, and when Ripley stabilizes its position, the 3-D-Objecter waits 0.5 seconds to ensure that the 2-D-Objecter's region report is stable, and then resumes 3-D processing. When the 3-D-Objecter processes a 2-D region set, it projects each region to a corresponding 3-D object location. Then, the projected objects are then placed into correspondence with existing objects in the 3-D mental model. To compare projected and existing objects, a modified version of (1) is used in which $d_p()$ measures 3-D Euclidean distance, and $d_s()$ measures size. The same alignment process as the 2-D-Objecter is used to align projected objects to existing objects in the mental model. If projected objects have no existing counterparts in the simulator, new objects are instantiated. Conversely, if an object exists in the mental model but no corresponding object has been projected based on visual evidence, then the object in the mental model is destroyed. There is no hysteresis function required in the 3-D-Objecter since all 2-D regions have already passed through a hysteresis function in the 2-D-Objecter.

E. Inferring Force Vectors From Vision

In the process of updating the position of moving objects, the 3-D-Objecter must infer the magnitude and direction of forces which lead to observed motions. Inference of force dynamics has been argued to be of fundamental importance in grounding verb meanings [40]. We now explain how forces are inferred from visual observation in the Objecter.

Consider a situation in which an object, such as a ball, is on the workspace and in view. Once the 2-D-Objecter has registered the corresponding region, it will relay the region to the 3-D-Objecter which will instantiate an object in the mental model. At this point, Ripley is aware of the ball. Now, assume the ball begins to slowly roll. Although the visual region corresponding to the ball will be displaced from one time step to the next, the 2-D-Objecter will generally determine the correspondence between regions over time steps and thus track the object. After the correspondence process has been run by the 3-D-Objecter, a displacement in positions between projected and existing objects in the simulator must be accounted for. This is where the force inference step takes place. A force proportional to the displacement and in the direction of the projected object is applied within ODE to the corresponding object. As the object accelerates (decelerates), the inferred forces will be increased (decreased) accordingly. To summarize, in the process of tracking objects, the Objecter also generates a constant stream of inferred forces acting on each object to account for their changes in velocity. These force vectors may be used to classify self-moving objects, and other aspects of force dynamics.

F. Generating Images Within the Mental Model

The mental model is integrated with a 3-D graphics rendering environment [41]. The 3-D environment may be rendered from an arbitrary viewpoint by positioning and orienting a synthetic camera and rendering the scene from the camera's perspective. Changes in placement of the synthetic camera are used to implement shifts in perspective without physically moving Ripley. Fig. 5 shows an example of a synthetic view of the situation also depicted in Figs. 3 and 4. Words with visually referential semantics (*blue*, *ball*, *left*, etc.) are grounded in terms of features extracted from these synthetic "mental images." As we shall see, we can ground spatial phrases such as *my left* as a combination of a shift of perspective combined with a visually grounded spatial model.

V. Situated Speech Understanding and Generation Grounded in Mental Imagery

The mental model and mental imagery provide Ripley with object permanence and imagined perspective shifts, enabling new forms of human-machine dialog. As a first exploration into its use, we have integrated the architecture into a dialog system that supports early forms of spoken dialog with Ripley. This integrated system consists of several components including a sensorimotor grounded lexicon, a speech recognition and robust parser, grounded semantic composition procedures, and visually driven language generation procedures. Although complete descriptions of these modules is beyond the scope of this paper, we briefly sketch salient aspects of each module so that the application of the mental model and imagery may be presented.

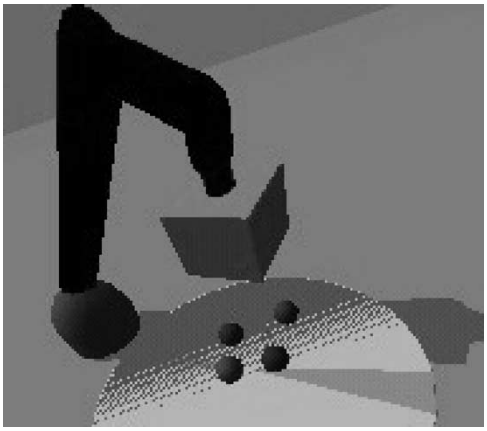

Fig. 5. By positioning a synthetic camera at the position approximating the human's viewpoint, Ripley is able to "visualize" the scene from the person's point of view, which includes a view of Ripley.

A. Grounded Lexicon

A central component of the system is a grounded lexicon that defines the meaning of words in terms of richly structured sensorimotor representations. In essence, these structures model the meaning of words in terms of their correspondences to percepts, actions, and affordances [42].

B. Verbs = Sensorimotor Networks (SNs)

The meaning of manipulation verbs (*lift*, *pick up*, *touch*) are grounded in SNs (a closely related approach can be found in [21]). SNs can be used to execute actions on the robot (in that sense, they may be thought of as plan fragments), but they also serve as a representational substrate for the semantics of verbs, and modifiers that are linked to verbs.

A SN is defined by a linked set of *perceptual conditions* and *motor primitives*. Fig. 6 shows the SN for *pickup*. Perceptual conditions are indicated by rectangles, motor primitives by circles. Verbs expect a single argument x, the patient of the verb.[2] The main execution path of this SN is a single alternating sequence of perceptual conditions and motor primitives. The *pickup* SN may be interpreted as: 1) ensure x is in view; 2) extend head until x is visually looming (recall that Ripley's cameras are mounted next to the gripper); 3) grasp with the gripper until the gripper touch sensors are activated; and finally 4) retract. Errors can be sensed at each perceptual condition. The default behavior on all errors is to retry the previous motor action once, and then give up. All SNs terminate in either a *success* or *failure* final state.

C. Modifiers = Sensorimotor Expectations

Modifiers such as color, shape, and weight are defined with respect to an underlying SN. Fig. 7 illustrates the representation of *heavy* and *light*. This structure captures the commonsense notion that something is heavy if it is difficult to lift. The

[2]In ongoing work, we are expanding our formalism to accept agents, instruments, and manner arguments.

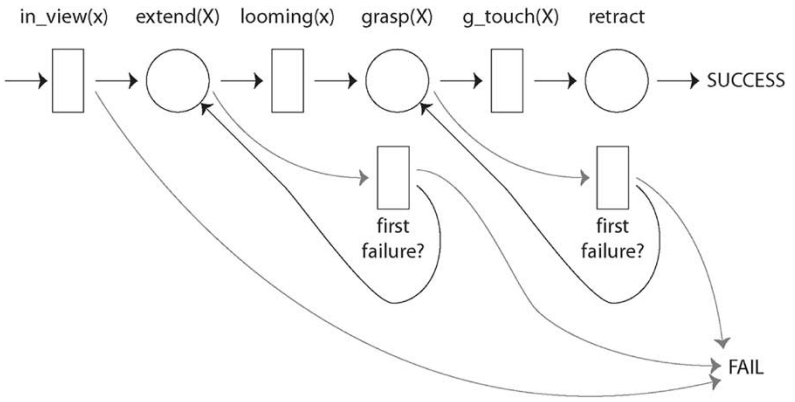

Fig. 6. A SN that encodes the semantics of *pickup*.

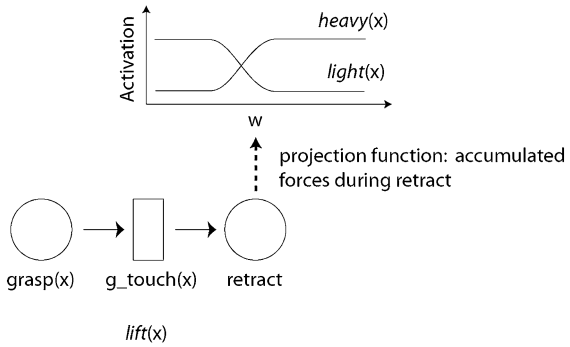

Fig. 7. The meaning of *heavy* and *light* are grounded in expected resistance measurements while lifting an object.

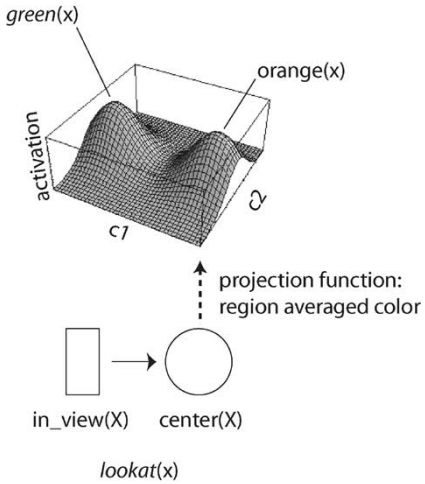

Fig. 8. The meaning of *green* and *orange* are grounded in expected distributions of context-normalized color space measured by looking at an object.

SN (bottom) grounds the meaning of *lift*. The dashed line indicates a *projection function* that projects the execution of an SN into a low dimensional feature space. In this case, the projection function accumulates joint forces during the execution of the *retract* motor primitive, effectively weighing the patient of *lift*. The meaning of *heavy* and *light* are grounded as distributions of expected values with respect to this projection of the underlying SN. These distributions are referred to as *activation functions*. To determine how well a word fits an object, the SN underlying that word must be executed and projected using the associated projection function. The activation function associated with the word is evaluated at the projected point to determine how well the word fits the object. Since activation functions are continuous, all scores are continuously graded.

Categorical distinctions (e.g., determining whether an object is blue or not, as a binary decision) are made using a simple voting mechanism. Within a feature space, the most activated function determines the category label of the object.

The grounding of color terms closely parallels weight terms (Fig. 8). In place of *lift*, color terms are defined in terms of the SN associated with *lookat*, which, when executed, causes Ripley to center the object x in the robot's visual field. The projection function computes the average value of color in all pixels of the visual region corresponding to the object. Color terms such as *green* and *orange* are defined as 2-D Gaussian distributions within this projected feature space.

Shape descriptors are grounded using histograms of local geometric feature, described in [43]. The histograms are generated using a projection function defined in terms of the same SN as color terms (*lookat*).

D. Spatial Relations and Perspective Shifting

To ground spatial words (e.g., *above*, *to the left of*) in our past work with 2-D virtual worlds (cf. [14]), we have used Regier's set of three spatial features [11], which take into account the relative shape and size of objects. The first feature is the angle (relative to the horizon) of the line connecting the centers of area of an object pair. The second feature is the shortest distance between the edges of the objects. The third feature measures the angle (relative to the horizon) of the line which connects the two most proximal points of the objects. Spatial relations such as *above* and *left of* are defined as Gaussian distributions in terms of these three features. To apply a spatial relation, two objects must be identified, the target and the landmark. The two-argument structure associated with spatial terms is encoded in the speech parser as described below.

Since Ripley's mental model is 3-D, we use projective transforms to capture 2-D views of the mental model (using synthetic vision). Regier's features are then computed on the 2-D image. In Regier's models, and our previous work, the perspective of the viewer has always remained fixed, assuming

a first person perspective. Using the mental model, the synthetic camera can be moved to any 3-D location and orientation. Using this perspective shift operation, the semantics of *my left* versus *your left* can be differentiated by using the word *my*, in this linguistic context, as a trigger for positioning the synthetic camera. Ripley's proprioceptive system guides the placement of the camera for first person perspectives, and the face-tracker driven human model enables shifting to the human's point of view.

E. Spatially Situated Speech Understanding

Using the SN and projection function representation, we have encoded a small vocabulary of words that cover verbs (*pickup*, *touch*, etc.), names of objects (*apple*, *beanbag*, *cup*, etc.), and terms for color, weight, and spatial relations. A speech recognizer, parser, and semantic composition system work together to convert commands into robot actions. Most aspects of the lexical structures are hand coded. Only the activation functions (e.g., color distributions associated with *green*, or weight distributions associated with *heavy*) are trained from examples using standard statistical estimation techniques.

Front end speech recognition is performed using a HMM-based decoder [44]. The single best word sequence is passed to a chart parser [45] which serves as the first step of a semantic composition procedure. The composition process is presented in detail in [46]. In brief, each lexical entry has a function interface that specifies how it performs semantic composition. Currently, the interface definition consists of the number and arrangement of arguments the entry is willing to accept. Semantic type mismatches are handled during composition rather than being enforced through the interface. Each entry can contain a *semantic composer* that encapsulates the actual function to combine this entry with other constituents during a parse.

The system is able to resolve the referent of utterances with multiple modifiers. To achieve this, virtual objects consisting of one or more actual objects are internally generated during semantic composition. Consider the spoken command, "Pick up the large green cup to the left of the blue plate." To resolve the reference of *large green cup*, the innermost term, *cup*, is first bound to objects in the robot's environment based on the visual shape models associated with the word. If multiple cups are found, they are grouped into a virtual object. This virtual object is then composed with the representation of *green*, which will threshold and sort the contents of the virtual object based on greenness, and pass along the new virtual object to *large*. The landmark phrase *blue plate* is processed in the same way, resulting in a second virtual object. The spatial phrase *to the left of* is used to find the best pair of objects, one drawn from each of the virtual objects. Finally, the best referent is passed as an argument to the *pickup* SN, which actually executes the action and picks up the target object.

The words *my*, *your*, *me*, and *you* are given special treatment when adjacent to spatial terms, each triggering an appropriate shift of visual perspective within Ripley's mental model (*in front of me*, *to your left*, etc.). Subsequent spatial terms are evaluated in the shifted frame of reference. In this way, mental imagery provides the grounding for deictic and intrinsic spatial language.

F. Results and Discussion

Ripley is able to interactively respond to a range of imperative spoken commands such as "Pick up the blue cup on your left" and "Hand me the ball to the right of the large green beanbag." In cases where the referent of the command appears ambiguous, Ripley uses a simple dialog strategy to request further descriptive terms. When an explicit spatial frame is not indicated through language, Ripley's default is to imagine the workspace from the user's point of view, thereby interpreting commands from a deictic frame of reference. Ripley is able to understand commands which specify single actions to be performed on single objects. More complex request sequences of actions or manipulation on multiple objects is currently beyond the scope of the system's grammar.

In principle, Ripley can understand reference to objects that are not in its camera's view due to the object permanence function of the mental model. Although the physical camera may not be directed toward a target object, the synthetic camera can be directed at any portion of the 3-D model of the scene to ground referring expressions. We have not yet implemented the procedures for controlling synthetic vision, but the representational capacity for performing this kind of language comprehension is in place.

Although this application of the mental model is too preliminary for formal performance analysis, three major sources of processing errors are apparent, each suggesting a direction for future work. First, the simplifying assumption of color-based foreground/background separation in the low-level visual analysis algorithms leads to significant limitations of the vision system. To address this, a more robust segmentation process based on contrast maps and depth imaging is being developed. Second, several parameters in the Objecter are manually set (distance tuning weights, confidence decay rates, etc.), leading to sub-optimal synchronization of the model to the robot's environment. Instead, these parameters can be automatically determined using machine learning techniques, once an annotated set of active vision data has been collected. Third, the speech recognizer occasionally produces errors. There are several ways to improve robustness of speech recognition. These include the use of acoustic confidence scores to reject poorly recognized words, and the integration of speech processing with contextual knowledge derived from other sensors (see [47] for first steps in this direction).

VI. Conclusions

Our vision is to create interactive robots that can engage in cooperative tasks with humans mediated by fluid, natural spoken conversation. To achieve this vision, the robots must have rich representations of the physical situations in which they are embedded. These representations must be coupled to the robot's physical senses so that it reflects reality, and provide appropriate interfaces for grounding natural language.

Motivated by these needs, we have developed a method for constructing and maintaining a physical model of a robot's environment based on active perceptual input. The mental model provides a representational medium that is suitable for

grounding the semantics of referring expressions. The mental model serves as the robot's dynamically constructed "cache" of the external world. Rather than tie the meaning of utterances to first-person perspective visual representations, the mental model provides an abstracted representational layer to interface with natural language semantics.

Understanding relativized spatial language is only one of numerous reasons for endowing Ripley with a mental model. Consider, for example, how Ripley should generate referring expressions to bring its human partner's attention to an object. Depending on the situation, objects in view for the robot may be occluded from the human's perspective. If a cup is sitting behind an obstacle, say a box, that prevents the human from seeing the cup, it would be ineffective for Ripley to refer to the cup as just *the cup*. Instead, by taking into account the human's viewpoint, Ripley can anticipate that the object will not be in view and instead say *the cup behind the box*. Ripley's mental model enables this kind of situated language use.

Perhaps one of the most intuitive views of word meaning is the referential theory: words get their meaning due to their correspondence to events, objects, properties, and relations in the world. Although many other critical aspects of meaning have been raised in the philosophy of language and mind, the referential aspect of words holds firm as a crucial part of any complete theory of meaning. The approach we have presented here enables Ripley to establish meaningful correspondence between words and world, enabling a central aspect of situated language understanding.

ACKNOWLEDGMENT

The authors express their thanks to P. Gorniak and N. Mukherjee for implementing Ripley's low-level image processing and speech parsing modules. The authors also thank the feedback from anonymous reviewers which led to significant improvement in the presentation of the material.

REFERENCES

[1] J. Barwise and J. Perry, *Situations and Attitudes*. Bradford, MA: MIT-Bradford, 1983.
[2] H. Clark, *Using Language*. Cambridge, MA: Cambridge Univ. Press, 1996.
[3] M. K. Brown, B. M. Buntschuh, and J. G. Wilpon, "SAM: A perceptive spoken language understanding robot," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, pp. 1390–1402, Nov./Dec. 1992.
[4] C. Crangle and P. Suppes, *Language and Learning for Robots*. Stanford, CA: CSLI Publications, 1994.
[5] P. McGuire, J. Fritsch, J. Steil, F. Roethling, G. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter, "Multi-modal human-machine communication for instructing robot grasping tasks," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, vol. 2, Aug. 2002, pp. 1082–1088.
[6] D. Perzanowski, A. Schultz, W. Adams, K. Wauchope, E. Marsh, and M. Bugajska, "Interbot: A multi-modal interface to mobile robots," in *Proc. Language Technologies 2001*. Pittsburgh, PA, 2001.
[7] W. Bluethmann, R. Ambrose, M. Diftler, S. Askew, E. Huber, M. Goza, F. Rehnmark, C. Lovchik, and D. Magruder, "Robonaut, A robot designed to work with humans in space," *Auton. Robots*, vol. 14, pp. 179–197, 2003.
[8] Z. Kazi, S. Chen, M. Beitler, D. Chester, and R. Foulds, "Grasping at straws: An intelligent multimodal assistive robot," in *Proc. Int. Conf. Rehabilitation Robotics*, Bath, U.K., 1997.
[9] J. Feldman, G. Lakoff, D. Bailey, S. Narayanan, T. Regier, and A. Stolcke, "Lzero: The first five years," *Artific. Intell. Rev.*, vol. 10, pp. 103–129, 1996.
[10] J. M. Lammens, "A computational model of color perception and color naming," Ph.D. dissertation, State Univ. New York, New York, 1994.
[11] T. Regier, *The Human Semantic Potential*. Cambridge, MA: MIT Press, 1996.
[12] T. Regier and L. Carlson, "Grounding spatial language in perception: An empirical and computational investigation," *J. Experim. Psych.*, vol. 130, no. 2, pp. 273–298, 2001.
[13] D. Roy, "Learning words from sights and sounds: A computational model," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, 1999.
[14] ——, "Learning visually-grounded words and syntax for a scene description task," *Comput. Speech Lang.*, vol. 16, no. 3, 2002.
[15] ——, "Grounded spoken language acquisition: Experiments in word learning," *IEEE Trans. Multimedia*, vol. 5, pp. 197–209, June 2003.
[16] J. Siskind, "Naive physics, event perception, lexical semantics, and language acquisition," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, 1992.
[17] D. Roy, "Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic," *J. Artific. Intell. Res.*, vol. 15, pp. 31–90, 2001.
[18] C. Yu, D. Ballard, and R. Aslin, "The role of embodied intention in early lexical acquisition," in *Proc. Cognitive Science Soc.*, Boston, MA, 2003.
[19] D. Bailey, "When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs," Ph.D. dissertation, EECS Dept., Univ. California at Berkeley, Berkeley, CA, 1997.
[20] K. Hsiao, N. Mavridis, and D. Roy, "Coupling perception and simulation: Steps toward conversational robotics," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Las Vegas, NV, 2003.
[21] S. Narayanan, "Karma: Knowledge-based active representations for metaphor and aspect," Ph.D. dissertation, Univ. California at Berkeley, Berkeley, CA, 1997.
[22] P. Johnson-Laird, *Mental Models: Toward a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Cambridge Univ. Press, 1983.
[23] G. Fauconnier, *Mental Spaces*. Cambridge, MA: MIT Press, 1985.
[24] The Stanford Encyclopedia of Philosophy [Online]. Available: http://plato.stanford.edu
[25] R. A. Finke, *Principles of Mental Imagery*. Cambridge, MA: MIT Press, 1989.
[26] P. McKellar, *Imagination and Thinking*. London, U.K.: Cohen & West, 1957.
[27] L. Barsalou, "Perceptual symbol systems," *Behav. Brain Sci.*, vol. 22, pp. 577–609, 1999.
[28] A. Glenberg, "What memory is for," *Behav. Brain Sci.*, vol. 20, pp. 1–19, 1997.
[29] J. Piaget and B. Inhelder, "The childs conception of space," in *The Essential Piaget*. New York: Basic Books, 1948, pp. 576–642.
[30] B. Tversky, "Structures of mental spaces: How people think about space," *Environ. Behav.*, vol. 35, pp. 66–80.
[31] M. Skubic, G. Chronis, P. Matsakis, and J. Keller, "Spatial relations for tactical robot navigation," in *Proc. SPIE, Unmanned Ground Vehicle Technology*, Orlando, FL, 2001.
[32] G. Miller and P. Johnson-Laird, *Language and Perception*. Cambridge, MA: Harvard Univ. Press, 1976.
[33] D. Roy, P. Gorniak, N. Mukherjee, and J. Juster, "A trainable spoken language understanding system for visual object selection," in *Proc. Int. Conf. Spoken Language Processing*, Denver, CO, 2002.
[34] F. Mussa-Ivaldi and E. Bizzi, "Motor learning through the combination of primitives," *Philos.Trans. R. Soc. London*, vol. 355, pp. 1755–1769, 2000.
[35] Open Source Computer Vision Library [Online]. Available: http://www.intel.com/research/mrl/research/opencv
[36] R. Smith. (2003) ODE: Open Dynamics Engine. [Online]. Available: http://q12.org/ode/
[37] F. Cao and B. Shepherd, "Mimic: A robot planning environment integrating real and simulated worlds," in *Proc. IEEE Int. Symp. Intelligent Control*, 1989, p. 459 464.
[38] W. J. Davis, "On-line simulation: Need and evolving research requirements," in *Handbook of Simulation: Principles, Methodology, Advances, Applications and Practice*, J. Banks, Ed. New York: Wiley, 1998.
[39] J. R. Surdu, "Connecting simulation to the mission operational environment," Ph.D. dissertation, Texas A&M, College Station, 2000.
[40] L. Talmy, *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press, 2000.
[41] OpenGL website [Online]. Available: www.opengl.org

[42] J. J. Gibson, *The Ecological Approach to Visual Perception*. Mahwah, NJ: Erlbaum, 1979.
[43] D. Roy, B. Schiele, and A. Pentland, "Learning audio-visual associations from sensory input," in *Proc. Int. Conf. Computer Vision Workshop on the Integration of Speech and Image Understanding*, Corfu, Greece, 1999.
[44] B. Yoder, "Spontaneous Speech Recognition Using Hidden Markov Models," Master's thesis, Massachusetts Inst. Technol., Cambridge, MA, 2001.
[45] J. Allen, *Natural Language Understanding*. New York: Benjamin Cummins, 1995.
[46] P. Gorniak and D. Roy, "Grounded semantic composition for visual scenes," *J. Artif. Intell. Res.*, to be published.
[47] N. Mukherjee and D. Roy, "A visual context-aware multimodal system for spoken language processing," in *Proc. Eurospeech*, Geneva, Switzerland, 2003.

Deb Roy received the Bachelor's degree in computer engineering from the University of Waterloo, Waterloo, ON, Canada, and the M.S. and Ph.D. degrees from Massachussetts Institute of Technology (MIT), Cambridge.

He is an Associate Professor of Media Arts and Sciences, AT&T Career Development Professor, and directs the Cognitive Machines Group, all at the MIT Media Lab.

Kai-Yuh Hsiao received the B.S. and M.Eng. degrees in electrical engineering and computer science in 1999 and 2001, respectively, from Massachussetts Institute of Technology (MIT), Cambridge, where he is currently pursuing the Ph.D. degree at the MIT Media Lab, working on grounding language in robots.

Nikolaos Mavridis received the B.S. degree in mathematical sciences from the U.K. Open University in 1998, the M.Eng. degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1999, and the M.S. degree in electrical engineering from the University of California at Los Angeles in 2000. He is currently puruing the Ph.D. degree at the Media Laboratory, Massachussetts Institute of Technology (MIT), Cambridge, concentrating in mental models and active sensing.