To Ask or To Sense?

Planning to Integrate Speech and Sensorimotor Acts

Nikolaos Mavridis and Haiwei Dong
New York University Abu Dhabi
P.O. Box 129188, Abu Dhabi, UAE
{nikolaos.mavridis; haiwei.dong}@nyu.edu

***Abstract*—For machines to converse with humans, they must at times resolve ambiguities. We are developing a conversational robot which is able to gather information about its world through sensory actions such as touch and active shifts of visual attention. The robot is also able to gain new information linguistically by asking its human partner questions. Each kind of action, sensing and speech, has associated costs and expected payoffs with respect to the robot's goals. Traditionally, question generation and sensory action planning have been treated as disjoint problems. However, for an agent to fluidly act and speak in the world, it must be able to integrate motor and speech acts in a single planning framework. We present a planning algorithm that treats both types of actions in a common framework. This algorithm enables a robot to integrate both kinds of action into coherent behavior, taking into account their costs and expected goal-oriented information-theoretic rewards. The algorithm's performance under various settings is evaluated and possible extensions are discussed.**

Keywords-attention, active perception, ambiguity resolution, discourse planning

I. INTRODUCTION

Consider the following scenario:

Figure 1. Human, robot, green can and red ball

A human and a robot are sitting across a table that holds two objects, a red ball and a green can. The human has selected one of the objects on the table, and asks the robot to give it to him: "Hand me the ball!"

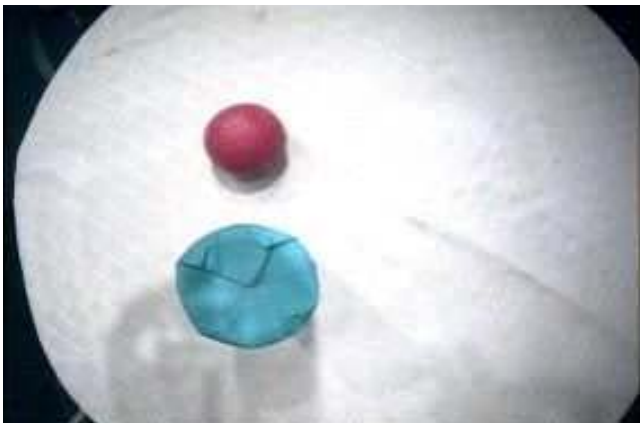

Figure 2. Robot's view of can and ball

The robot is viewing the objects from above (the robot has video cameras mounted next to its gripper). From this perspective, both objects appear circular, and thus the robot cannot clearly distinguish between the ball and the can without further information. What should the robot's next move be? Should it move its head to a different position in order to get a view from a different perspective, and thus hope- fully resolve the ambiguity? Or perhaps the robot should ask its human communication partner a clarifying question ("Do you mean the red one?"). To make a decision between these actions, the robot needs a coherent framework for com- paring expected costs and payoffs of motor acts and speech acts. Moving wears down gears and heats up motors, but may reveal new visual information that resolves the ambiguity. Asking questions may annoy the human, but might lead to an unambiguous description. Moving motors might be a wasted effort (both objects might turn out to be balls). Asking a question might not lead to disambiguation. This example captures the gist of the problem we wish to address. This paper describes a first step in this direction by considering just one kind of speech act (clarifying questions about object descriptions) and one kind of sensorimotor act (measuring perceptual properties). Ultimately, we seek a unified framework for planning a variety of motor and speech acts.

Properties such as color and size would seem to be known as soon as a person (or robot) looks at an object. In general, however, measuring and storing information is in general expensive and thus should be performed only when needed. It is well known that people do not encode seemingly obvious

978-1-4673-2015-3/12/$31.00 ©2012 IEEE

information about visual scenes. For example, you might easily recall who you met at a party a week ago but not re- call the color of a small triangle that decorated his tie, even if it was within your field of view for a couple of minutes. And this might not have just been forgotten, but instead might not have been noticed in the first place. Studies in change blindness [1, 2] suggest that attention to visual properties depends on function. Information that is relevant to the task at hand is more likely to be attended to and remembered. Studies of visual attention in humans also provide insight into processes that we are trying to model. Patterns of gaze trajectories (cf. [3]) determine which parts of the scene will be most probably noticed. The gaze trajectory seems to be highly scene-specific, as well as task-specific. Furthermore, briefly fixating over an object doesn't guarantee that all of its observable properties will be noticed; the task at hand seems to bias which properties are most probably noticed. For example, in a size-matching task, intricacies of texture might pass unnoticed. This seems to be caused not primarily due to limitations of short-term visual memory, but most probably due to the activation of specific computational resources that carry out the processing tasks associated with the estimation of various visual properties (see also [4]). Furthermore, such a decomposition of visual processing into attentional shifts and primitive estimation operations is reminiscent of Shimon Ullman's visual routines proposal [5] and Rao's language of attention [6].

The above example provides the main motivation for the simple model that is presented here. We will describe a robot planning algorithm that resolves ambiguous verbal descriptions of objects. This problem is motivated by our goal of developing conversational robots that can perform collaborative tasks with human partners based on situated speech understanding and generation. We have concentrated on the problem of deciding between sensing physical properties of an object versus asking clarifying questions to resolve the referent of an object description. We will assume that each primitive sensing action aims at providing information about a specific property of a specific object. In our model, the properties considered will be color, size, and weight. On the other hand, speech actions will have the form of disambiguating questions, querying the human about properties of the intended referent, such as: "Is it red?" or "How heavy is it?".

Given sufficient information, the robot will eventually find the object that the human had in mind, and thus will be able to execute spoken commands. In essence, the referent is resolved through a sequence of (sensory question/world answer) and/or (speech question/human answer) pairs, and many such possible pairs are possible. Thus the question the robot must address is: given a predefined set of allow- able moves, how does one select what the next action should be? Traditionally, sensorimotor action generation has been accounted for in the field of "active perception", as for ex- ample in [7]. On the other hand, question generation is considered a sub-topic of discourse planning. Here, we treat both in a common framework.

In the selection of the next move, two often conflicting goals should be taken into account. First, we need a move that will provide useful information towards our ultimate goal - resolving the referent. We are thus motivated to develop a solution in an information theoretic framework. Inspite of their information value, actions also have costs which might arise for numerous reasons: computational cycles, time and energy spent, annoying the human partner with questions, and so forth. The robot must select sequences of actions that minimize long term expected costs while achieving its goals. Given a way to calculate expected benefit and pre- defined costs, we can create an algorithm that chooses the next move, in either an optimal or suboptimal but satisfactory manner (in the sense of "satisficing" [8], and thus utility theory becomes relevant. Examples of information and utility theoretic models for the case of language planning include [9], and for the case of vision [7, 10].

The planning algorithm presented here is by no means optimal, as it contains numerous simplifying assumptions and approximations. However, this was not only acceptable but also quite welcome, as our goal was to produce behaviors that would subjectively appear as human-like.

Sensory actions and spoken questions have indeed different domains of inquiry: the former query the physical state of the world, while the latter query the mental state of the human partner. But in essence they are both just sub-types of generalized sensing acts, seeking information from the world about the world, and are thus treated uniformally here.

Initially the model of the situation will be presented, followed by a section describing the procedure for the choice of the next move. Then, experimental results will be given, for quantitative simulations as well as real-world tests on a robot. Finally, possible extensions of the model along various axis will be discussed, followed by a conclusion.

II. THE MODEL

The model has been developed for Ripley, a robot that is able to manipulate objects on a table top [15]. We consider a situation in which there are n objects on the table, and the human has selected one of them and wishes to communicate that choice to the robot. Initially, the robot encodes only the presence of the n objects but does not measure any of its properties. We model the representation of objects by both human and robot as a bundle of perceptual properties: size, color and weight. The robot has at its disposal a set of actions that it may take to uncover unknown property values of the intended object:

Speech action Q1: Ask "What <attribute> is it?" (e.g., "What color is it?", "How heavy is it?", "What size is it?")

Speech action Q2: Ask "Is it <attribute category>?" (e.g., "Is it green?", "Is it heavy", etc.)

Sensory action A1: Measure <attribute> of object i (e.g., "Measure color of object 1")

Sensory action A2: Measure <attribute> of all objects (e.g., "Measure size of all objects")

Although the sensory measurements return continuous values, we assume that they are further quantized to their equivalent verbal categories. A predetermined and mutually

agreed upon set of such categories is assumed to exist for both human and robot, and no negotiation of meaning needs to take place. In our case:

Size belongs to {small, medium, large} Color belongs to {red, green, blue, cyan, magenta, black} Weight belongs to {light, heavy}

At each step, the robot retains internally a description of the state of the world as best known, as well as a history of actions taken so far. The state consists of attribute category probability distributions for the n + 1 objects, namely the n objects and the intended referent. Thus, for example, at some step in the process of disambiguation we might have the state depicted in Fig. 3, which is interpreted as follows: The robot already knows that the intended referent is light and small, and that it is not red. Both objects have been measured as being small, their weights are unknown so far, and object 1 is green while the color of object 2 has not been measured yet. Such a situation might have arisen in the following scenario:

The initial human utterance was "Hand me the small light object". The robot measured the size of object 1, and then object 2. The robot did not choose to measure weights at this point, due to the high cost of such an action (the robot would have to grasp and lift the object to determine its weight). Instead, the robot chose to measure the size of object one, and it was small. Then, it chose to measure the size of object two, and it was small, too. Thus, both objects had equal probability of being the referent at this stage. Thus, the robot then chose to ask: "Is it Red?" and the human answered: "No". The robot measured the color of object one, and it was green, and so on.

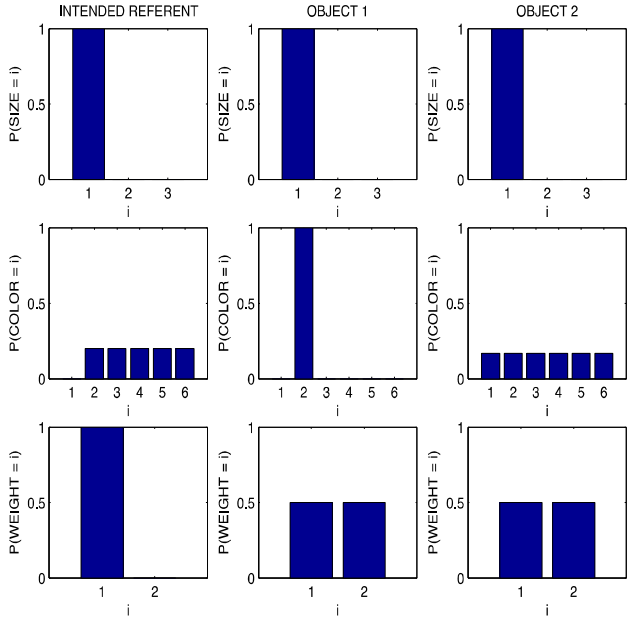

Figure 3. A state of the world

Notice that in the model described here, probability distributions are discrete, and we have complete confidence in measurements or answers. We also assume that the at- tribute dimensions are independent, and that the robot starts with non-informative priors. Both of these assumptions can be relaxed in the future.

Another basic assumption of the model is that the intended referent can be uniquely discovered given the action set available and a cooperative user (providing answers to speech actions) as well as the state of the situation. For the purposes of this initial model, we have disallowed object-specific disambiguating questions (the robot cannot ask: "Is it object 1?"). This means that objects chosen by the human must be unique in terms of perceptual properties. Thus, if an object is small, green and light, no other small, green and light object can exist on the table. However, two other small, red and light objects might exist simultaneously. In the future, spatial relations and object-specific questions can naturally complement the current model.

III. CHOOSING THE NEXT MOVE

Given the robot's knowledge of the state of the world so far, we seek to calculate the probability of the intended referent being a specific object, i.e. $P(I = O_k)$ for every object. This is a discrete distribution over the objects, and is approximated as follows: (see appendix for derivation)

$$P(I = O_k) = \frac{\prod_j \sum_i P_I(at_j = i) P_{O_k}(at_j = i)}{\sum_k \prod_j \sum_i P_I(at_j = i) P_{O_k}(at_j = i)} \quad (1)$$

where $P_I(at_j = i)$ is the probability attribute j has value i for the Intended referent and $P_{O_k}(at_j = i)$ is the probability attribute j has value i for Object k.

Our goal is to choose the next action such that it satisfies the possibly contradictory requirements of lowering the expected uncertainty of $P(I = O_k)$ after the results of the action, and also achieving that at a low cost. The current costs of the actions are assumed fixed. Looking forward, these values would eventually be provided by a central "re- source pricing" module in the robot, which values the cost of allocating computational and sensory motor resources given the current demand and limitations. Given no such mechanism, the real significance of the cost values lies in the preference ordering that they impose upon actions of equivalent uncertainty reduction, and their power to override uncertainty reduction in significance when large cost differences exist. This will become clear later, when we discuss how we address fusing benefit versus cost. Thus, if the costs are not expressed in units of any physical meaning, their ordering is what matters, except in the case of huge differences.

The expected uncertainty reduction resulting from a possible move is calculated through a one-step look-ahead procedure, which accounts for all permissible outcomes of the action under consideration. Outcomes that are contradictory to previous evidence or the uniqueness of solution assumption are discarded. One-step lookahead is suboptimal compared to multi-step, but is preferred initially for its simplicity and the production of human-like behavior. We calculate the average entropy after all possible outcomes of each action, as follows:

$$\tilde{H}(act_i) = E\left(-\sum_k P(I = O_k)\log P(I = O_k)\right) \quad (2)$$

where the expectation operator $E(\cdot)$ is taken over all permissable outcomes of action i, giving the expected value of their resulting $P(I = O_k)$ distributions.

Now the problem of fusing benefit versus cost is addressed. We would like to capture the following intuitions in a single cost function:

1) For equal information rewards, minimum costs should dictate the preference□

2) For small and medium cost differences, information gains should be dominant

3) For large cost differences, minimum cost might override the ordering imposed by information gains alone

The following functional form was thus chosen, satisfying the above requirements:

$$Value(act_i) = -\tilde{H}(act_i) - \frac{e^{cost(act_i)}}{k} \quad (3)$$

where:□

k is a freely chosen fixed constant
$cost(act_i)$ is the predetermined cost of action i, and□
$Value(act_i)$ is the resulting value of action i after taking into account information benefit versus action cost ($\max = 0$)

After being selected, the move with greatest reward not belonging to the set of moves already executed is selected for execution. The process continues until the intended referent is successfully resolved.

In summary, each action not yet taken is considered in turn, and the effects of each of its permissible outcomes are evaluated in terms of information gain. Then, the expected benefit is fused with the cost through the above formalism, giving the value of each action. The action of maximal value is selected and carried out. The state of the world is then changed according to the results. Finally, this action selection / information gathering / state update cycle is repeated until $P(O_k = I)$ approaches 1.0 for a specific object, indicating that the intended referent was successfully resolved.

IV. RESULTS

Two types of test trials were run: quantitative evaluation trials with a simulated environment and human, as well as real world trials with a conversational robot and human partner.

For the quantitative trials, random configurations of the world and the intended referent were generated by sampling the underlying property distributions, and rejecting configurations in case the uniqueness of intended referent assumption was violated. The algorithm's performance was quantified in terms of the mean values and variances of the number of moves and the total cost required for successful referent resolution. Two baselines for result comparison were used, namely totally random choice of moves without repetition, and choice of moves without taking uncertainty reduction into account. Results for various settings are presented below:

TABLE I. AVERAGE NUMBER OF MOVES AND COST

Mode	$\mu_{n_{moves}}$	μ_{cost}
entropy and cost	8.25	3.41
cost only	10.45	4.68
random	13.9	5.99

Using the proposed method, we get an improvement of 25 percent in average total cost, and 12 percent in average number of moves, given over the baseline of cost-only and non-entropy-based action selection, and an even greater improvement given over the trivial strategy of random action selection. It is easy to see that the exact improvement on the average total cost depends on the inhomogeneity of the cost values across actions. Given an adequately inhomogeneous distribution of costs, we would expect a similar figure. On the other hand, the average number of moves criterion carries a more general quantitative significance.

The above results correspond to randomized trials with four objects. Cost values were arbitrarily adjusted in order to elicit preference of sensory actions over questions, and to avoid the time and energy consuming procedure of weighing objects. As discussed in the previous section, if the cost values do not correspond to any physically meaningful units, then their actual values matter only as far as they induce an altered preference order on the actions, given their information value. The costs that were used, corresponding to the action preference order described, are shown in Table 2.

TABLE II. AVERAGE NUMBER OF MOVES AND COST

Type	**Action**	**Cost**
A1	Get size of one	0.17
A1	Get color of one	0.23
Q2	Is it size ?	0.33
Q2	Is it color ?	0.44
Q1	What size ?	2.85
Q1	What color ?	3.44
A2	Get sizes of all	5.91
A2	Get colors of all	5.86
A1	Get weight of one	7.41
Q2	Is it weight ?	7.52
Q1	What weight ?	7.63
A2	Get weights of all	8.00

The algorithm was also evaluated in the real world on Ripley, a conversational robot [15]. The model was integrated with a speech synthesis module, a speech recognizer, the mental model of the robot for querying apparent object properties, and the robot action interface. The sensing actions

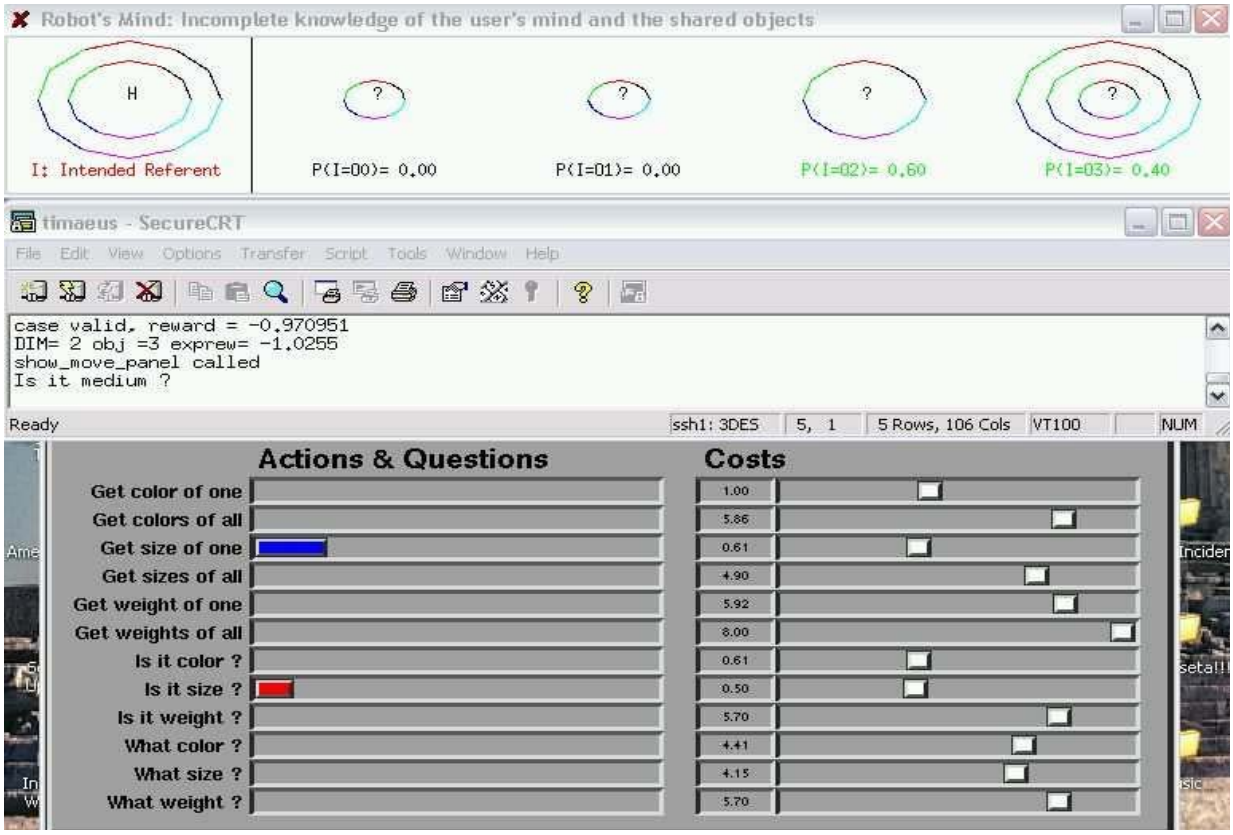

Figure 4. Resover running on Ripley.

corresponding to size and color measurements were carried out by the robot's vision system. Weight sensing was carried out through an active measurement procedure: the robot picks up the object, swings it in a predetermined manner, and the force applied by the motors is measured and converted to a weight value.

The algorithm was also evaluated in the real world on Ripley, a conversational robot [15]. The model was integrated with a speech synthesis module, a speech recognizer, the mental model of the robot for querying apparent object properties, and the robot action interface. The sensing actions corresponding to size and color measurements were carried out by the robot's vision system. Weight sensing was carried out through an active measurement procedure: the robot picks up the object, swings it in a predetermined manner, and the force applied by the motors is measured and converted to a weight value.

A screenshot of resolver running on Ripley is shown in Fig. 4. The cost panel is visible at the lower part of the screen. The arbitrary cost values are set through the movement of the sliders, as a central resource pricing center is not currently available on Ripley. The red and blue color bars correspond to the total cost spent in specific actions taken already. At the middle part of the screen, the last question that was spoken through the robot's speech synthesizer is visible: "Is it medium?". Finally, at the upper part of the screen is a visualization of the state of the robot's knowledge of the intended referent and the four objects. The superposition of possible attribute values is depicted. For example, the robot knows that the intended referent can either be medium or large, but not small. Thus, only two concentric circles of medium and large radii are shown, but not the small one. Also, the robot knows that the intended referent might have any allowable color, and so the circles have arcs colored with red, green, blue, magenta, cyan and black. The interpretation of the state of the robot's knowledge on the four objects is similar. Finally, the probability values of $P(I=O_k)$ are shown below the four objects. At this stage, we are certain that the intended referent can neither be object 0 nor object 1. It is most probably object 2, with probability 0.6, but can also turn out to be object 3, with probability 0.4.

V. Future Directions

The model presented here is a simple starting point. The problem of interleaving speech acts with motor acts in a conversational robot is of course an extremely challenging problem. In this section, we present some research threads that we plan to explore as extensions of the current model.

In the derivation of Eq. (1) presented in the appendix, we have assumed attribute independence as well as flat non-informative priors. However, this is fortunately far from true in the sensory world, where strong natural modes exist[13]. For example, a banana-shaped object is most likely yellow. If we

had used empirical priors and models of dependence, then some observations or answers would be much more informative. Thus, the average number of moves or cost required for successful referent resolution would decrease noticeably.

Given the possibility of inconsistent answers and measurements, one can try to devise correction and backtracking schemes. However, this can automatically be taken care of by assuming partial (and not total) confidence in sensory or human answers. Such a situation could be caused by imperfect hearing or sensing, or an unreliable human or sensory world. Thus, the update rules for $P_I\left(at_j = i\right)$ and $P_{O_k}\left(at_j = i\right)$ presented in the appendix would have to be- come softer. Then, one would expect that certain matching can never be achieved; and thus, satisficing criteria must be decided on an acceptable certainty of match, as well as numerous associated heuristics. Also, given the above situation of imperfect hearing or sensing, and through the interdependence between objects and the intended referent, there is more to be potentially gained: One could actually allow correcting what is heard based on what is seen and vice versa, in a similar manner as in [14].

Another direction would be the introduction of a continuous distribution at the sensory measurement level (instead of the discrete $P_{O_k}\left(at_j = i\right)$). The choice of the form of the distribution is free. Also, one can impose a full bayesian treatment instead of the simple ad-hoc approximations shown in the appendix.

Furthermore, other primitive questions/actions or molecular combinations of the ones given might be explored. An obvious first choice is a saccade primitive action, moving the fixation point from object to object, with a cost increasing with distance. Such an action would have to be combined with the existing primitive sensing actions and treated as a molecule, in our case of single step-ahead planning.

Also, without relaxing any of the above assumptions but by doing full-horizon (and not only single step) prediction with no approximations, one can try to solve the problem optimally (by brute force, dynamic programming, or some heuristic). This would also create an upper bound on performance in terms of minimum attainable average cost, which would be useful as a further baseline for performance evaluation.

Finally, it is worth stressing here again what the intended purpose of this model is. This is not an attempt towards an optimal solution; we just want something simple, at a behavioral complexity level that appears cognitively plausible, that will produce behavior that could subjectively be judged as reasonably human-like. Thus, another more remote direction of extension would be to use the model presented here as a predictive model for human performance, given suitably designed experiments with eye-tracking etc. Then, its similarity to human performance could also be assessed more objectively.

VI. Conclusion

A conversational robot must connect language, perception, and action at numerous levels from lexical semantics to planning. In this paper, we have presented an initial model of planning to resolve ambiguous verbal descriptions by interleaving question generation with active sensing. The model has been implemented and integrated into a real-time conversational robot. This work constitutes a step towards our long term vision of situated human-machine communication.

Acknowledgment

The authors would like to thank Deb Roy for his helpful comments, as well as Kai-Yuh Hsiao for his help regarding motor system modifications for the robot trials.

References

[1] H. Intraub, "The representation of visual scenes," *Trends in Cognitive Sciences,* vol. 1, no. 6, pp. 217-222, September 1997.

[2] D. J. Simons and D. T. Levin, "Change blindness," *Trends in Cognitive Sciences,* vol. 1, no. 7, pp. 261-267, October 1997.

[3] J. Triesch, D. H. Ballard, M. M. Hayhoe and B. T. Sullivan, "What you see is what you need," *Journal of Vision,* vol. 3, no. 1, pp. 86-94, January 2003.

[4] A. Treisman, "The pereption of features and objects," in *Attention: Selection, awareness and control*, A. Baddeley and L. Weiskrantz, Eds., Oxford, Clarendon Press University, 1993, pp. 5-35.

[5] S. Ullman, "Visual routines," *Cognition,* vol. 18, no. 1-3, pp. 97-159, December 1984.

[6] S. Rao, "Visual routines and attention," PhD Dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2002.

[7] R. Pito, "A Solution too the next best view problem for automated surface acquisition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 21, no. 10, pp. 1016-1030, October 1999.

[8] H. A. Simon, The Sciences of the Artificial, Cambridge, MA: MIT Press, 1969.

[9] T. Paek and E. Horvitz, "Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems," in *AAAI Fall Sumposium on Psychological Models of Communication in Collaborative Systems*, Cape Cod, MA, 1999.

[10] P. Whaite and F. P. Ferrie, "Autonomous exploration: driven by uncertainty," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, no. 3, pp. 193-205, March 1997.

[11] M. A. Arbib, "The mirror system, imitation, and the evolution of language," in *Imitation in Animals and Artifacts*, C. Nehaniv and K. Dautenhahn, Eds., Cambridge, MA: MIT Press, 2000.

[12] L. Steels, "Mirror neurons and the action theory of language origins," in *International conference on Architectures of the Mind, Architectures of the Brain*, Vatican City, Vatican, 2000.

[13] W. Richards, Natural Computation, Cambridge, MA: MIT Press, 1988.

[14] D. K. Roy and N. Mukherjee, "Visual context driven semantic priming of speech recognition and understanding," *Computer Speech and Language.*

[15] N. Mavridis and D. Roy, "Grounded situation models for robots: where words and percepts meet," in IEEE International Conference on Intelligent Robots and Systems, pp. 4690-4697, 2006.

Appendix
Algorithm Details

An approximate simplified discrete model with uniform priors, independence assumptions, and full confidence in human / sensory answers is used. However, it is easily

modifiable to a continuous version and towards the relaxation of the above assumptions. At each step, the knowledge of the situation so far is represented through the distributions of the attribute categories of the intended referent and the objects:

$P_I\left(at_j = i\right)$: probability of attribute j having the value i for the intended referent.

$P_{O_k}\left(at_j = i\right)$: probability of attribute j having the value i for object k.

Initially, these are uniform distributions. After each human/sensory answer, they are updated, through multiplication with an answer-specific term and renormalization to unit sum:

Speech action Q1: Ask "What <attribute j> is it?" Human answer: "It is <attribute category c>"

$$P_I\left(at_j = i\right) \text{ multiplied by } \begin{cases} 0, & i \neq c \\ 1, & i = c \end{cases}$$

Speech action Q2: Ask "Is it <attribute j category c>?" Human answer: "Yes"

$$P_I\left(at_j = i\right) \text{ multiplied by } \begin{cases} 0, & i \neq c \\ 1, & i = c \end{cases}$$

Human answer: "No"

$$P_I\left(at_j = i\right) \text{ multiplied by } \begin{cases} 1, & i \neq c \\ 0, & i = c \end{cases}$$

Sensory action A1: Measure <attribute j > of object i Sensory answer: "category c"

$$P_{O_k}\left(at_j = i\right) \text{ multiplied by } \begin{cases} 0, & i \neq c \\ 1, & i = c \end{cases}$$

Sensory action A2: Measure <attribute j > of all objects Same as A1, taking each object in turn

Also, after new results from sensory actions, $P_I\left(at_j = i\right)$ is multiplied with the product of the relevant $P_{O_k}\left(at_j = i\right)$.

Towards the goal of calculating $P\left(I = O_k\right)$ as given in Eq. (1), the following steps are taken:

$$P\left(at_j \text{ being equal for } I \text{ and } O_k\right) = \sum_i P_I\left(at_j = i\right) P_{O_k}\left(at_j = i\right)$$

(after the simplifying assumption that P_I and P_{O_k} are independent, incorrect but tolerated).

$$P\left(\text{all attributes equal for } I \text{ and } O_k\right) = \prod_j P\left(at_j \text{ being equal for } I \text{ and } O_k\right)$$

(after the independence across attributes assumption and commuting sums over products).