# Task Coordination and Assistive Opportunity Detection via Social Interaction in Collaborative Human-Robot Tasks

Aaron St. Clair and Maja J. Matarić
Interaction Lab, Computer Science Department
University of Southern California, Los Angeles, CA
{astclair, mataric}@usc.edu

#### **ABSTRACT**

In most environments, task collaboration requires efficient, flexible communication between collaborators. In the case of tasks involving human-robot collaboration, the robot must effectively convey and interpret communicative actions about the current and intended state of the task environment and coordinate its behavior with those of its collaborators, human and otherwise. A framework for collaborative communication should not only allow the robot to reason about its actions in relation to those of others but should also support reasoning about the robot's own intentions and those ascribed to it by the collaborators. We present such a framework, inspired by Theory of Mind and making use of perspective taking, and show how it could be used to support several collaborative functions, including detection opportunities to assist.

**KEYWORDS:** collaborative systems, human-robot interaction, cognitive modeling, socially assistive.

#### 1. INTRODUCTION

In collaborative task settings, people use a combination of explicit social signaling and implicit situational awareness to communicate with each other in order to achieve coordinated, goal-driven action. Robots are becoming increasingly capable of conducting high-level tasks in environments, including manipulating real-world everyday objects and navigating cluttered spaces. Humanrobot collaboration necessitates principled methods of integrating human perception and modeling and task control to allow a robot to collaborate effectively on a task, through learning or direct control specification. In collaborative contexts, it is assumed that the interacting agents share a joint goal; this provides the robot with an opportunity for the use of perspective-taking to evaluate

others through the use of its own control structure. With human collaborators, the robot should also be capable of formulating and responding to natural communication, including gestures, proxemics, and simple speech commands, to achieve coordination and clarify intent. Although people tend to use similar forms of communication across different tasks, the dynamics are dependent on the environment, user tendencies, and the task itself. To enable a robot to communicate naturally, we propose combining task actions and communicative actions into a unified, flexible framework that also enables the identification of assistive opportunities.

## 2. BACKGROUND

In order to develop robots capable of natural, social coordination behavior with humans, we first consider human collaboration and the insights it offers.

#### 2.1. Human Collaboration

It has been documented that people have a tendency to adapt both their speech and actions in response to the person they are interacting with, so as to be salient and sensible to a collaborating partner, especially in circumstances involving work objects in the environment and frames of reference [1]. Collaborators align their linguistic representations of the environment, allowing for more effective communicative behavior with their partner(s). This alignment is achieved via a process in which local environmental representations, i.e., specific speech and gesture, are implicitly adopted and propagated to global representations via a priming mechanism [2]. The same priming mechanism is similarly used to achieve lexical and syntactic alignment, resulting in a consistent vocabulary and shared environmental representation at the task level, for corresponding communication. Researchers disagree over how deeply people model their interaction partners and the role those models play in language and gesture production. Some argue that a speaker's model of addressees plays a central role in production [3]; others suggest it functions as a late stage corrective mechanism for tailoring speech [4]; and some suggest a hybrid approach is used, depending on context [5]. Additionally, people rely heavily on speech for coordination, but it has been shown that shared visual information can result in more efficient, less verbose utterances [6]. This suggests that collaboration can be accomplished without relying exclusively on natural language processing, particularly in collocated scenarios.

#### 2.2. Human-Robot Collaboration

Prior work on human-machine collaboration includes approaches from human-computer (HCI) and humanrobot interaction (HRI), as well as from cognitive science and linguistics. An extensive body of work from Grosz et al. on top-down deliberative approaches for modeling is aimed at establishing and maintaining alignment, to assure coherent discourse, and constructing shared plans for collaboration [7,8]. This approach has been applied in HCI to intelligently alter the user interface and reduce the total amount of communication required to complete certain tasks [9]. There has also been extensive work on intent recognition relying on perspective taking or Theory of Mind-inspired models to allow a robot to recognize a person's intentional behavior through observation. This approach has been employed to learn the rules of a series of games by clustering estimated intents into roles that a robot can assume [10], as well as to recognize and generate intentional face-to-face meeting initiation behavior [11], and to recognize helping and hindering social behavior via an MDP formalization utilizing inverse planning [12]. Work by Breazeal [13] and Hoffman [14] demonstrated the ability of a robot to learn simple tasks through human tutelage and collaborate effectively via turn-taking or biased pre-emptive action. Related work in multi-robot systems generally separates collaborative communication from task function through a repeated process of team partitioning and reformation for achievement goals [15] or through conditional behavior switching for maintenance goals [16].

#### 3. APPROACH

Our work is aimed at integrating social communication with task control, learned or otherwise, to support coordination in collaborative environments across different tasks that the robot already knows how to perform independently. The specific context of collaboration allows for the simplifying assumption that participants are working together to accomplish a shared set of goals. This enables the robot to make use of its own task controller or planner to evaluate the actions of others. To accomplish this, it must first actively model its collaborators, possess a manipulable representation of the

task, and be able to evaluate the environment from other perspectives. These capabilities enable the robot to predict and coordinate with others, to identify assistive opportunities, and to detect whether its plans are aligned with others. Assistive conditions in the context of collaborative activity can be classified across two dimensions: 1) according to their motivating stimuli, i.e., whether they are aimed at error recovery or preemptive aid, and 2) whether the robot takes a communicative or a task action in response. For example, a robot may detect the occurrence of an error and it might offer a collaborator encouragement through social interaction (communicative recovery) or it might attempt to fix the error itself (taskbased recovery). Similarly, some assistive actions such as reinforcement, clarification, or preemptive completion, do not depend on the prior occurrence of an

## 3.1. Theory of Mind

In order to effectively coordinate a robot's actions with those of its human collaborator, the robot must be able to accurately estimate the human's planned actions from context or from explicit communication. Analogously, the robot must be able to effectively convey its planned actions clearly to a human. This ability to attribute mental state to others and use it to plan and predict behavior is called Theory of Mind (ToM) and has been extensively used for various capabilities with autonomous robots. We propose a ToM-inspired model in which the robot contains estimates of its own state, the state of collaborators, and those collaborators' estimates of the robot's state. These states contain information relevant to the task including a world model and a partial task allocation i.e. assignments of various agents to sub-tasks. Previous work has demonstrated the viability of similar frameworks to model and learn from human activity at various levels of perceptual abstraction. At the task-level, it is generally assumed that the environmental dynamics are accurately detectable by the robot and the modeling relies on some notion of symbolic state [12,13], while other work focuses on bottom-up learning from raw or annotated sensor input [11,14,17,18]. Our work is focused on enabling natural collaborative communication in realistic collaborative task settings devoid of a strict dialog structure, such as turn taking. Thus, the task-level symbolic states will be augmented to allow the robot to formulate collaborative communication and infer intention from human social signals from perceptual features, such as head direction and deictic gesturing, extracted from on-board sensing.

## 3.2. Task Representation

Since our primary focus lies in modeling collaborative communication and detecting assistive opportunities, we assume a symbolic, STRIPs-like representation of the task in terms of pre- and post-conditions, and robust, though not perfect, sensing of environmental dynamics. Other work has demonstrated that task structure can be learned via a variety of mechanisms including through demonstration [19,20] and human tutelage [21,22] and thus our focus will not be on learning the specific task structure but rather on using an existing task controller to plan the robot's actions and evaluate those of its collaborators, while adding a social communication planning layer. A symbolic task representation is required to allow the robot to predict others' actions, formulate partial plans for each interacting partner, and coordinate its actions to achieve progress toward the goal. It also allows for integration of social communication via a separate controller, by evaluating the relative value of conducting a task action or performing communication actions, such as deictic references or other gestures.

## 3.3. Perspective-Taking

In order for the robot to make use of this state and task representation, it must be able to evaluate the environment from multiple points of view. Existing approaches to perspective taking generally rely on transforming sensor data to a local reference frame and planning under the estimated visibility constraints for another agent [13,23]. We will make use of a similar approach wherein the robot will assume a limited visual field of view for each participant and a distance-based efficiency metric. The estimated field of view of each participant will be inferred from tracked pose information. consisting of global position and body and head orientation. This allows filtering of salient environmental targets that are not in the given agent's field of view. The task controller, which maps a given state to a subtask, will then be used to obtain a prediction of possible next states given the position and visibility constraints for each other agent. Each agent's model of the robot's next state will be constructed similarly, using a filtered environment map.

In most real-world environments there are manageable subsets of objects that will play a role during performance of the task. Thus, constructing an environmental map involves tracking a subset of key objects throughout the collaboration. This could be accomplished with a tailored detection system for a specific task domain or could be learned via activity discovery, wherein the system observes someone performing the task while tracking salient features and infers what is important for each segment of the activity [24]. In cluttered or partially observable environments, additional robots could be used to provide better monitoring of the task space and human collaborators. The performance of the perspective-taking approach will be evaluated separately from the rest of the system to ensure that it is reasonably accurate at

predicting the actions of others. For tasks in which action is heavily tied to proximity, such as herding, foraging, or constrained object manipulation, we anticipate this approach to work well. For complex tasks with many action choices or where action is decoupled from global position, this approach may need to be augmented to learn the task preferences of a given individual, in order to account for personal differences such as performing similar tasks consecutively despite distance constraints.

#### 3.4. Uncertainty

Our goal is to develop systems that will operate in real-world environments with people and thus must handle imperfect sensing, partial observability, and uncertainty. The framework above will be used as the basis for a probabilistic reasoning approach in order to deal with sensor noise and uncertainty in state estimations and action outcomes. All the state maintained by the robot, including its own, those of others, and the perceived environmental dynamics, will be represented as distributions. This structure will scale with the number of collaborators and the number of states in the task, but will be tractable for everyday manipulation or maintenance tasks where the number of states is manageable.

## 3.5. Collaborative Reasoning

Given this framework, the robot can consider the consequences of its actions, both in terms of manipulating the environment and conveying information to the collaborative partners. Modeling sub-goal dynamics allows the robot to convey information implicitly by using its task selection or spatial positioning to project the information. In addition, the robot can explicitly communicate information by selecting actions such as gesturing and vocalization. Finally, this framework allows the robot to compare the estimates of itself with those of others so as to detect and handle discrepancies, identify assistive situations, and recognize agents using a different plan than itself, potentially allowing it to adapt to or instruct others. Next, we describe how the above framework could be used for simple action coordination and detection of these and other special circumstances during the course of a collaborative activity with a human.

# 3.5.1. Coordinating Task Action

The simplest function of a collaborative system is to coordinate actions and minimize conflict with other agents while exploiting opportunities for parallel activity. This can be accomplished using the above framework by identifying parallelizable subtasks, tasks with met preconditions that have not been completed, and inhibiting the task controller from doing tasks that are likely to be performed by other agents.

Similarly, we can integrate communication actions such as pointing, gestures, or narration, by augmenting the set of subtasks or task actions with communicative actions that could be performed at any time. The collaboration monitor can then compare the estimated intentions of the robot from the various participants point of view with its actual planned intention. In ambiguous or uncertain cases the robot can then take an explicit communicative action to attempt to reinforce its intentions and increase fluency with its teammates. Deciding the type, timing, and frequency of such feedback depends to some extent on the task, although certain gestures such as deictic pointing are used extensively in collocated contexts. Gestural patterns could be learned from human-human interactions; initially, we will be employing a heuristic approach by selecting from a small set of well-understood gestures or narrations.

# 3.5.2. Detecting Assistive Opportunities

In most tasks, multiple situations are likely to arise wherein the robot could potentially offer assistance to a collaborator. These circumstances include struggling or failing to accomplish an intended subtask, not knowing what to do next, or employing an incorrect strategy or one different from that used by the robot. All of these cases require the robot to monitor the other participants' intentions and outcomes over time. In the above framework, this process consists of saving the estimated intentions of each participant at each step in the task and using learned or heuristics-based models of failure to recognize when things go wrong. The amount of past state required depends on the failure detection time-scale. Detecting success and failure of an atomic task could be accomplished using only a single recorded state while detecting successful but incorrect actions could require a complete model of every possible task state. Related work has shown nontrivial gameplay trajectory modeling to be tractable [25]. It is also possible that systemic differences in overall task strategy could occur for tasks with multiple viable strategies. In those cases, the robot's perspectivetaking capability would fail, resulting in it failing to predict intentions and potentially undoing progress made by other agents. Since the robot can no longer perform the task, it must resort to learning the collaborator's strategy, teaching the collaborator its own or teleoperation.

#### 3.5.3. Command Hierarchies

It is likely that different people will have different preferences for the role of the robot and its level of autonomy during collaboration. In some cases, it may be necessary for the robot to assess its place in the command hierarchy and adjust its level of autonomy. Extending the framework described above to model desires in addition to intentions and providing an on-line feedback mechanism for collaborators would allow the robot to adjust its action selection accordingly and thus improve performance.

#### 4. VALIDATION AND OUTCOMES

The described approach will be validated on a challenging cooperative task in a dynamic environment involving humans and robot(s), in which effective communication, rather than a one-time partitioning of responsibility, is required to achieve a collaborative goal. Such tasks could include kitting or object retrieval and construction tasks where the actions of one agent are tightly-coupled to the possible actions of others. Full details of the validation task and system will be provided in a later paper. We will demonstrate the ability of the framework to capture the explicit communication required to predict human intentions, identify a variety of assistive opportunities and complete the task. The generalization of the framework across tasks and users will be demonstrated on a secondary task, through a series of human subjects experiments aimed at comparing the system with a nonspeaking human confederate and a robot that does not perform user modeling. Since the proposed approach will be targeted at human-robot interaction and may rely on detailed sensory information about the human such as head direction estimates, it is unlikely to scale for scenarios with many people or where the number of humans is much greater than the number of robots. In addition, since the robot(s) will be using the same onboard sensing for tracking people, monitoring environmental dynamics, and performing task actions, a view prioritization algorithm may be necessary.

#### 5. FUTURE WORK

We have described a method for modeling collaborative task-achieving behavior that we are in the process of developing and validating. The specific form and timing of the individual feedback gestures and the robot responses for the various assistive opportunities, for instance, merits additional study to verify that the robot clearly conveys its intentions. Further research into the best methods of perceiving humans is also warranted, to verify that the extracted social features extracted, such as location and head orientation, are suitable for monitoring typical human feedback. Finally, some tasks, such as those that employ exclusively reactive control, are not amenable to collaboration with this particular modeling method, since they do not provide symbolic output upon which planning can be performed. Our future work includes a rigorous investigation of the representational limits of the system for these various classes of tasks.

# **ACKNOWLEDGEMENTS**

This work was supported by National Science Foundation (NSF) grants CNS- 0709296, IIS-0803565, and IIS-0713697 and ONR MURI grant N00014-09-1-1031

#### **REFERENCES**

- [1] S. Whittaker, "Things to talk about when talking about things," *Human–Computer Interaction*, vol. 18, no. 1, pp. 149–170, 2003.
- [2] M. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, no. 02, pp. 169–190, 2004.
- [3] C. Lockridge and S. Brennan, "Addressees' needs influence speakers' early syntactic choices," *Psychonomic Bulletin & Review*, vol. 9, no. 3, p. 550, 2002.
- [4] B. Keysar, D. Barr, J. Balin, and J. Brauner, "Taking perspective in conversation: The role of mutual knowledge in comprehension," *Psychological Science*, vol. 11, no. 1, p. 32, 2000.
- [5] J. Hanna and M. Tanenhaus, "Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements," *Cognitive Science*, vol. 28, no. 1, pp. 105–115, 2004.
- [6] S. Fussell, R. Kraut, and J. Siegel, "Coordination of communication: Effects of shared visual context on collaborative work," ACM Conf. on Computer Supported Cooperative Work, p. 30, ACM, 2000.
- [7] B. Grosz and C. Sidner, "Attention, intentions, and the structure of discourse," *Computational Linguistics*, vol. 12, no. 3, pp. 175–204, 1986.
- [8] B. Grosz and S. Kraus, "The Evolution of SharedPlans," Foundations of Rational Agency, vol. 14, pp. 227–262, 1999.
- [9] N. Lesh, C. Rich, and C. Sidner, "Using plan recognition in human-computer collaboration," *International Conference* on *User modeling*, pp. 23–32, Springer-Verlag New York, Inc., 1999.
- [10] C. Crick and B. Scassellati, "Inferring narrative and intention from playground games," *IEEE International Conference on Development and Learning*, pp. 13–18, August 2008.
- [11] R. Kelley, C. King, A. Tavakkoli, M. Nicolescu, M. Nicolescu, and G. Bebis, "An architecture for understanding intent using a novel hidden markov formulation," *International Journal of Humanoid Robotics, Special Issue on Cognitive Humanoid Robots*, vol. 5, no. 2, pp. 1–22, 2008.
- [12] T. Ullman, C. Baker, O. Macindoe, O. Evans, N. Goodman, and J. Tenenbaum, "Help or hinder: Bayesian models of social goal inference," *Advances in Neural Information Processing Systems*, vol. 22, 2010.

- [13] C. Breazeal, J. Gray, and M. Berlin, "An embodied cognition approach to mindreading skills for socially intelligent robots," *International Journal of Robotics Research*, vol. 28, p. 656, May 2009.
- [14] G. Hoffman and C. Breazeal, "Effects of anticipatory perceptual simulation on practiced human-robot tasks," *Autonomous Robots*, vol. 28, pp. 403–423, May 2010.
- [15] M. Tambe and W. Zhang, "Towards flexible teamwork in persistent teams: Extended report," *Autonomous Agents* and *Multi-Agent Systems*, vol. 3, no. 2, pp. 159–183, 2000.
- [16] G. Kaminka, A. Yakir, D. Erusalimchik, and N. Cohen-Nov, "Towards collaborative task and team maintenance," *Int'l Conf. on Autonomous Agents and Multiagent Systems*, p. 73, ACM, 2007.
- [17] B. Scassellati, "Investigating models of social development using a humanoid robot," *Int'l Joint Conf. on Neural Networks*, vol. 4, pp. 2704–2709 vol.4, Jul. 2003.
- [18] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, no. 1, pp. 13–24, 2002.
- [19] M. N. Nicolescu, M.J. Mataric, "Natural methods for robot task learning: Instructive demonstrations, generalization and practice," *Int'l Conf. on Autonomous Agents and Multiagent Systems*. Melbourne, Australia, July 2003.
- [20] N. Koenig, L. Takayama, and M. J. Mataric', "Learning from demonstration: Communication and policy generation," *International Symposium on Experimental Robotics*, 2010.
- [21] A. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, no. 6-7, pp. 716–737, 2008.
- [22] C. Breazeal, G. Hoffman, and A. Lockerd, "Teaching and working with robots as a collaboration," *International Conference on Autonomous Agents and Multiagent Systems*, vol. 3, pp. 1030–1037, 2004.
- [23] J. Trafton, N. Cassimatis, M. Bugajska, D. Brock, F. Mintz, and A. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 35, no. 4, pp. 460–470, 2005.
- [24] D. Kirsh, "The intelligent use of space," *Artificial Intelligence*, vol. 73, no. 1-2, pp. 31–68, 1995.
- [25] D. L. Roberts, M. Nelson, C. L. Isbell, M. Mateas, and M. L. Littman, "Targeting specific distributions of trajectories in MDPs," *Twenty First National Conference on Artificial Intelligence*, 2006.