Task Learning through Visual Demonstration and Situated Dialogue

Changsong Liu and **Joyce Y. Chai**
Department of Computer Science and Engineering
Michigan State University, East Lansing, USA

Nishant Shukla and **Song-Chun Zhu**
Center for Vision, Cognition, Learning and Autonomy
University of California, Los Angeles, USA

Abstract

To enable effective collaborations between humans and cognitive robots, it is important for robots to continuously acquire task knowledge from human partners. To address this issue, we are currently developing a framework that supports task learning through visual demonstration and natural language dialogue. One core component of this framework is the integration of language and vision that is driven by dialogue for task knowledge learning. This paper describes our on-going effort, particularly, grounded task learning through joint processing of video and dialogue using And-Or-Graphs (AOG).

Introduction

As a new generation of social robots emerges into our daily life, techniques that enable robots to learn task-specific knowledge from human teachers have become increasingly important. In contrast to previous approaches based on Learning from Demonstration (Chernova and Thomaz 2014) and Learning by Instruction (She et al. 2014), we are currently developing a framework that enables task learning through simultaneous visual demonstration and situated dialogue. Supported by our framework, robots can acquire and learn grounded task representations by watching humans perform the task and by communicating with humans through dialogue. The long-term goal is to enable intelligent robots that learn from and collaborate with human partners in a life-long circumstance.

A key element in our framework is And-Or-Graph (AOG) (Tu et al. 2014; Xiong et al. 2016), which embodies the expressiveness of context sensitive grammars and probabilistic reasoning of graphical models. We use AOG to build a rich representation (i.e., STC-AOG) of the Spatial, Temporal, and Causal knowledge about the real world and the task. In addition, we are also designing an AOG-based schema (i.e., CI-AOG) to model and interpret the communicative intents between an agent and its human partner. These expressive and deep representations then allow a robot and a human to efficiently and effectively establish and increment their common ground (Clark 1996) in learning real-world tasks.

This paper provides an overview of the AOG-based framework and uses an example to illustrate our on-going work on joint task learning from visual demonstration and situated dialogue.

Representations

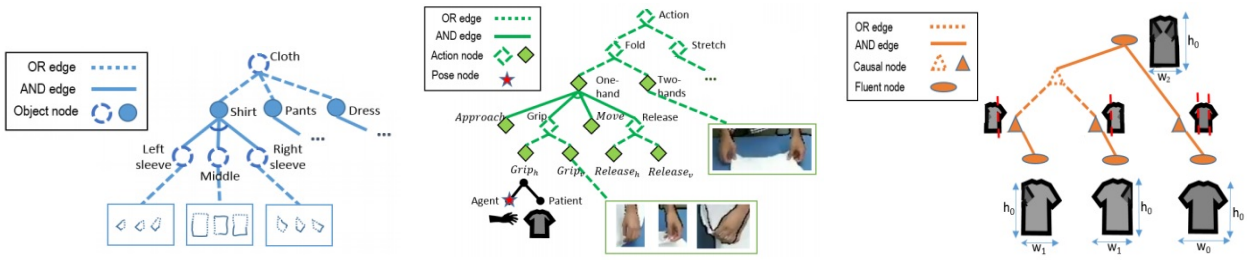
STC-AOG

An *And-Or-Graph* (*AOG*) (Tu et al. 2014) is an extension of a constituency grammar used in Natural Language Processing. It is often visualized as a tree structure consisting of two types of nodes, i.e., *And-node* and *Or-node*. An And-node represents the configuration of a set of sub-entities to form a composite entity; An Or-node represents the set of alternative compositional configurations of an entity. Using this general representation, three important types of task knowledge can be modeled:

- Spatial And-Or Graph (S-AOG) models the spatial decompositions of objects and scenes.
- Temporal And-Or Graph (T-AOG) models the temporal decompositions of events to sub-events and atomic actions.
- Causal And-Or Graph (C-AOG) models the causal decompositions of events and fluent changes.

Figure 1 illustrates an example of the S-/T-/C- AOG representation for cloth-folding tasks, which captures the spatial, temporal, and causal knowledge of the domain. Robots can then utilize this rich knowledge representation to understand, communicate, and perform task-oriented actions. Based on this knowledge representation framework, Xiong et al. (2016) has developed a statistical learning mechanism that automatically learns the parameters (e.g., the branching probabilities of Or-Nodes) of S-/T-/C-AOGs from a set of human demonstration videos. Furthermore, methods for learning the structures of different types of AOG have also been studied in previous work (e.g., Pei et al. 2013; Fire and Zhu 2013).

The basic idea of learning AOG-based task knowledge is to treat each demonstration as a specific instance, or a so-called "parse graph", which is generated by selecting one of the alternative configurations at each Or-node of an AOG model (see Tu et al. (2014) for details). Given a series of demonstrations represented as parse graphs, the structures and parameters of the underlying AOG model then can be learned using statistical learning techniques.



(a) S-AOG (b) T-AOG (c) C-AOG

Figure 1: An example of the S-/T-/C- AOG for a cloth-folding domain.

CI-AOG

Since AOG in essence can be viewed as a stochastic grammar machinery, and has been shown powerful in parsing the hierarchical structure of goal-driven events (Pei et al. 2013), we propose to use the same mechanism for analyzing the intentional structure of knowledge transferring dialogues.

For this purpose, we first construct an AOG, which we call the "Communicative Intent" AOG (CI-AOG) here, to describe how the intentional structure of such dialogues could possibly unfold. Our CI-AOG is similar to the T-AOG or "event grammar" as we illustrated earlier, where an Or-node captures different possibilities and an And-node captures sequential events, and the terminal nodes represent the basic actions (i.e., dialogue acts) that one can perform in a dialogue.

To illustrated the idea, we have manually crafted a (partial) CI-AOG that can be used to analyze the intentional structure of a task teaching dialogue as shown in Figure 2. We composed this CI-AOG based on "situated learning" literature (Lave and Wenger 1991; Herrington and Oliver 1995) to model how the teacher's and the learner's intents interact in a mixed-initiative dialogue. For example, we capture in this CI-AOG the common intentions in situated learning, such as *articulation* (the leaner articulates what is being understood regarding the current situation), *reflection* (the learner reflects what has been learned), and *assessment* (the teacher provides feedback to the learner's reflections or articulations).

Furthermore, the CI-AOG is also used to capture the unique characteristics of dialogue, including turn-taking, initiatives, and collaborative dynamics (Clark 1996; Clark and Schaefer 1989). To capture the turn-taking dynamics in dialogue, each node in CI-AOG is assigned a role (i.e., who the speaker is). This is illustrated in Figure 2 by assigning different colors to the nodes (i.e., orange nodes represent the learner and blue nodes represent the teacher). Therefore, an And-node in CI-AOG not only represents the temporal order of its children nodes, but also captures who takes the initiative of the sub-dialogue and how the turn-taking switches between the learner and the teacher.

The expressiveness of the AOG language also allows us to capture the collaborative dynamics studied in the discourse analysis literature (e.g., Clark and Schaefer (1989)). For example, as illustrated in the left bottom part of Figure 2, after the learner requests the teacher for teaching an alternative

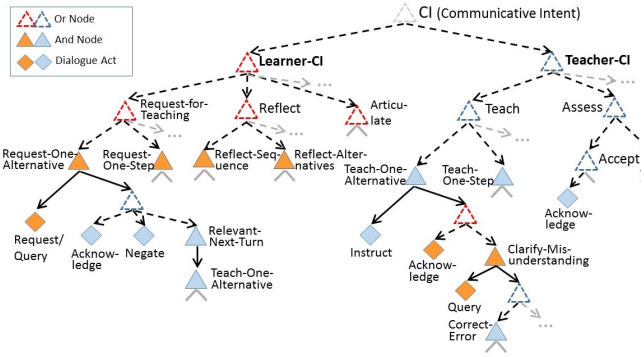


Figure 2: An example of Communicative Intent AOG (CI-AOG).

way of doing a task (i.e., the *Request-One-Alternative* node), the teacher should respond an explicit acknowledgement, or a negation, or directly teach an alternative without explicit acknowledging (the "relevant-next-turn" behavior).

Suppose a CI-AOG has already been constructed, it then can be used for "parsing" the underlying intentional structure of an ongoing dialogue. This is similar to previous work (Pei et al. 2013) that used a Top-Down parsing algorithm to analyze the hierarchical structure of goal-driven events from an observed sequence of atomic actions. Figure 3 further illustrates a parse graph on the underlying intentional structure of the following example dialogue.

Example dialogue of a robot learning to fold a t-shirt:

R1: Could you teach me how to fold the t-shirt?
H1: Sure.
H2: First, you fold the right sleeve towards the middle of the t-shirt.
R2: I saw you approached the right sleeve, grasped there, then moved to a position, and released your hand there.
R3: Is that position the middle of the t-shirt?
H3: Yes, that position is in the middle of the t-shirt.
R4: OK, what is the next step?
H4: Next, you fold the left sleeve to the middle.
R5: OK.
R6: This is what I have learned: first fold the right sleeve to the middle, and then fold the left sleeve to the middle.
H5: You can also fold the left sleeve first, and then the right sleeve.
R7: I see.
R8: What is next?
. . . . . .

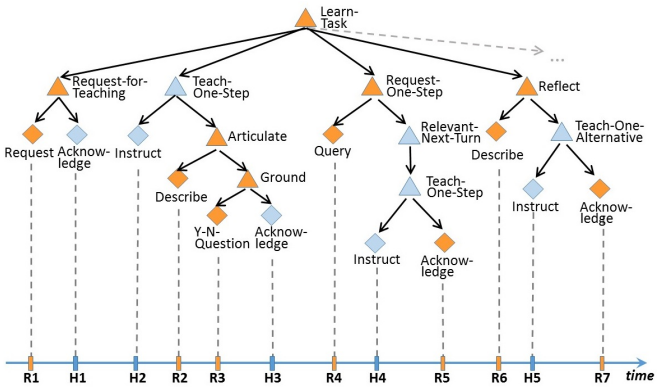


Figure 3: The CI parse graph for the given example dialogue.

As illustrated in Figure 3, the overall intent of this dialogue is for the robot to learn a sequential task. It is initiated by the robot's request for teaching (R1), followed by the human's explicit acknowledgement (H1). The following subdialogue is then led by the human's intent of teaching the robot the first step with an instruction (H2). Following that, the robot articulates what it understands about the current situation (R2), and tries to map the unknown concept "middle" to a physical position in the visual context (the question asked in R3, with an intent of what we call "*ground*"). The human's positive response (H3) confirms the robot's understanding, and also closes the subroutine of teaching the first step. The dialogue routine then rolls back to a higher-level of the intent hierarchy, where the robot moves on with its intent of learning the next step (R4). In R6, after two consecutive steps have been learned, the robot issues a reflection on what has been learned so far, which triggers human's following intent to teach an alternative order (H5).

Now we have introduced different types of AOG as the fundamental representations of the physical world, task knowledge, and dialogue dynamics. Next we turn our focus to discussing how we utilize these representations to build learning agents under a unified framework.

Learning from Situated Dialogue

Natural language and dialogue can play an important role in learning task knowledge from a human. Language provides a key source of information to gear the learned knowledge towards how humans conceptualize and communicate about situations and tasks. Such "human-oriented" knowledge is very necessary for facilitating human-robot communication and collaboration (for example, Lemon, Gruenstein, and Peters (2002)).

Furthermore, dialogue provides an expedited way to learn task knowledge. This can be demonstrated by our earlier example of learning how to fold a t-shirt. After the robot reflected (in R6) the just learned two steps (i.e., *fold-right-sleeve* and *fold-left-sleeve*), the human further taught that the order of the two steps could be switched and it would result into the same status of performing the task (H5). With our AOG-based representation, the robot can add this new knowledge by directly modifying the high-level structure of the STC-AOG (i.e., create new temporal and causal Or-Nodes to represent this alternative sequence of actions and fluent changes). Using language makes it much easier to communicate such high-level knowledge (Figure 4 illustrates the STC-AOG representation that can be learned thereafter).

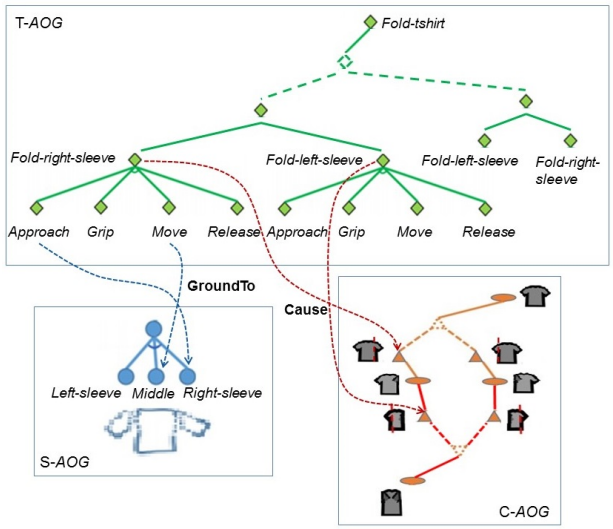


Figure 4: The STC-AOG representation of task knowledge that can be learned from the previous example dialogue of learning to fold a t-shirt. Note that the S-/T-/C- components are not independent from each other. The interplay between them provides an integrated representation of the task knowledge.

We thus propose an AOG-based framework to enable robot learning task knowledge from natural language and visual demonstration simultaneously. Supported by this framework, the robot can also proactively engage in human's teaching through dialogue, and gradually accumulate and refine its knowledge. One key advantage of our proposed framework is to provide a unified view of modeling the joint and dynamic task learning process. Besides, since we use AOG as a common representation basis, different components of our model can be stored and accessed using the same format (e.g., graph database), and be processed by the same set of algorithms. It thus can greatly ease the burden of building complex AI agents.

Figure 5 illustrates the basic ideas of our task learning system. It mainly consists of three tightly connected components that are all based on AOG representation and processing:

- *Language and Vision Understanding* processes the visual context into a "Vision Parse Graph" (*V-PG*) and the linguistic context into a "Language Parse Graph" (*L-PG*), and fuses them together into a "Joint Parse Graph" (*Joint-PG*) for a deep and accurate understanding of the current situation. A previous work (Tu et al. 2014) has employed the same AOG-based representations for joint text and video parsing in the question-answering domain. The processing in our component here resembles that work.

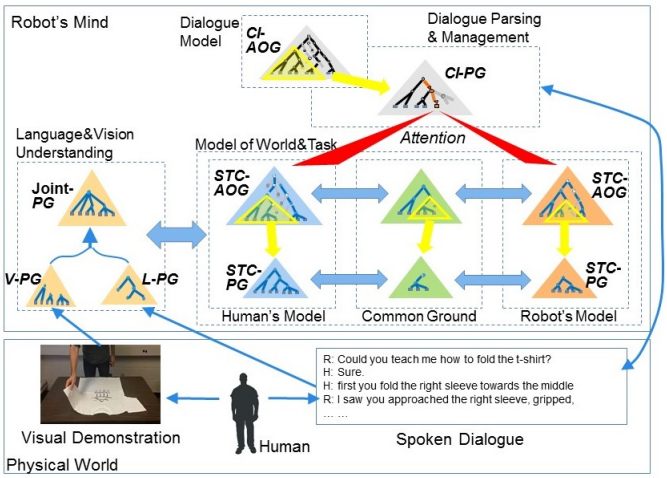


Figure 5: Illustration of our AOG-based framework for supporting robot learning from situated dialogue.

However the linguistic content of a dialogue could require more sophisticated approaches than those for handling monologues, and our goal is to learn generalizable task knowledge rather than just understand one situation.

- *World and Task Model* manages the representation and acquisition of knowledge of the physical world and tasks. As introduced earlier, we use STC-AOG to represent general knowledge about the world and the tasks, while a specific situation (i.e., a Joint Parse Graph) is represented as an instantiation (or sub-graph) of the STC-AOG. Motivated by the Common Ground theory (Clark 1996), our agent maintains three copies of models. One is the human's model of the world and knowledge, which is inferred from the joint parsing of language and vision. One is the agent's own model, and the third one is their shared/matched understanding of the situation and knowledge of the task (i.e., their *common ground*). In future work, we will further extend these models towards modeling the "Theory of Mind" in human-robot collaboration.
- *Dialogue Modeling and Management* uses CI-AOG to model and analyze the intentional structure of the task learning dialogue, and to facilitate the agent's decision making in knowledge acquisition and dialogue engagement. Our design of the situated dialogue agent also resembles the classical theory on discourse modeling (Grosz and Sidner 1986). I.e., the *intentional structure* is captured by a CI- Parse Graph (CI-PG) in our dialogue management component. The *linguistic structure* in our case has been extended to the joint model of the linguistic and visual contexts (captured as STC- Parse Graphs), and the shared knowledge (captured as STC-AOG). The *attentional state* is captured by linking each node in the CI-PG to a specific node/edge in the situation or knowledge representation graphs.

As the dialogue and demonstration unfold, the agent dynamically updates its intent, situation, and knowledge graphs. Each component can utilize the information from others through the interconnections between their graph representations. Based on this unified framework, sophisticated learning agents can become easier to be designed and built.

Conclusion

This paper provides a brief overview of our on-going investigation on integrating language, vision, and situated dialogue for robot tasking learning based on And-Or-Graphs (AOG). In particular, through an example, it demonstrates how language and dialogue can be used to augment visual demonstration by incorporating higher-level knowledge. Here we use cloth-folding as an example, but the same framework can be extended to other types of task learning. We are currently in the process of implementing the end-to-end system and plan to collect realistic data to evaluate our approach.

Acknowledgment

This work was supported by a DARPA SIMPLEX grant N66001-15-C-4035.

References

Chernova, S., and Thomaz, A. L. 2014. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 8(3):1–121.

Clark, H. H., and Schaefer, E. F. 1989. Contributing to discourse. *Cognitive science* 13(2):259–294.

Clark, H. H. 1996. *Using language*. Cambridge university press.

Fire, A. S., and Zhu, S.-C. 2013. Learning perceptual causality from video. In *AAAI Workshop: Learning Rich Representations from Low-Level Sensors*.

Grosz, B. J., and Sidner, C. L. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics* 12(3):175–204.

Herrington, J., and Oliver, R. 1995. Critical characteristics of situated learning: Implications for the instructional design of multimedia.

Lave, J., and Wenger, E. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge university press.

Lemon, O.; Gruenstein, A.; and Peters, S. 2002. Collaborative activities and multi-tasking in dialogue systems: Towards natural dialogue with robots. *TAL. Traitement automatique des langues* 43(2):131–154.

Pei, M.; Si, Z.; Yao, B. Z.; and Zhu, S.-C. 2013. Learning and parsing video events with goal and intent prediction. *Computer Vision and Image Understanding* 117(10):1369–1383.

She, L.; Yang, S.; Cheng, Y.; Jia, Y.; Chai, J. Y.; and Xi, N. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 89.

Tu, K.; Meng, M.; Lee, M. W.; Choe, T. E.; and Zhu, S.-C. 2014. Joint video and text parsing for understanding events and answering queries. *MultiMedia, IEEE* 21(2):42–70.

Xiong, C.; Shukla, N.; Xiong, W.; and Zhu, S.-C. Robot learning with a spatial, temporal, and causal and-or graph. Submitted to ICRA 2016.