A Unified Framework for Human-Robot Knowledge Transfer

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

Abstract

Robots capable of growing knowledge and learning new tasks is of demanding interest. We formalize knowledge transfer in human-robot interactions, and establish a testing framework for it. As a proof of concept, we implement a robot system that learns in real-time from human demonstrations.

Introduction

Transferring knowledge is a vital skill between humans for efficiently learning a new concept. Recently, skill acquisition and representation have become some of the core challenges in achieving robots capable of learning through human demonstration.

In a perfect system, a human demonstrator can teach a robot a new task by using natural language and physical actions. The knowledge can then be transferred back to another human, or further to another robot. The implications of effective human to robot knowledge transfer include the compelling possibility of a robot taking the role of a teacher, guiding humans in various tasks.

The technical difficulty in achieving a robot implementation of this caliber involves both an expressive knowledge structure and a real-time system for non-stop learning and inference. Most machine learning systems bifurcate learning and inference into separate steps. The learning step is usually trained off-line, resulting in an inference module only as powerful as the one-time learned model (Xiao et al. 2014). As a consequence, learning a new model from scratch becomes time-consuming and counterproductive to real-time systems. Moreover, most modern intelligent systems, such as Convolutional Neural Networks (CNN), learn a large number of parameters which do not clearly map to explicit spatial, temporal, or causal concepts.

We propose a real-time unified learning and inference framework for knowledge acquisition, representation, and transfer. Knowledge is represented in a Spatial, Temporal, and Causal And-Or Graph (STC-AoG) hierarchical network (Tu et al. 2014), which can be thought of as a stochastic grammar. The STC-AoG encapsulates the hierarchical compositional structures of physical objects, logical deductions, and instance-based actions. Our knowledge graph manipulation framework enables learning to be a continuous on-line process that occurs alongside inference. We view a robot as a knowledge database, where humans may deposit and withdraw skills. These skills can be used by both humans and robots alike.

As a proof of concept, we teach an industrial robot how to fold clothes (Figure 1). The robot watches a human demonstrator and learns in real-time. To test the faithfulness of the human-robot knowledge transfer, we propose an evaluation procedure called the Knowledge Transfer Test. Our experiments demonstrate that our proposed framework can adequately transfer knowledge to and from a robot. Furthermore, we illustrate our system's interactive learning capabilities that are backed by a Bayesian formulation.

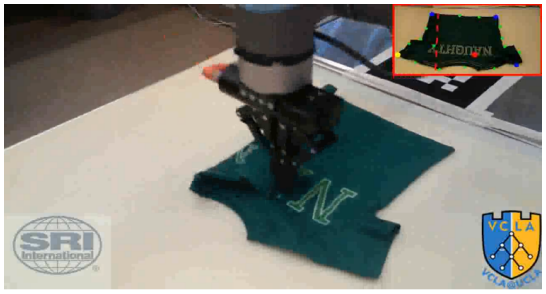

Figure 1: The robot performing a cloth folding task after learning from human demonstration.

Related Works

We extend the learning of And-Or grammars and semantics from video (Si et al. 2011) to an interactive real-time robotics platform with a natural communication interface between humans. The And-Or data structure has also been used in learning a visually grounded storyline model from labeled videos (Gupta et al. 2009); however, our system requires no labeled data, and evokes a richer segmentation of spatial, temporal, and causal concepts for more tractable queries. Miller et al. establish high standards for a cloth-folding robot, but our focus is instead on novel learning, knowledge representation, and knowledge transfer. The action-planning inference system in our STC-AoG

*Corresponding author

data structure resembles closest to a Planning Graph (Blum and Furst 1997), which is essentially an STC-AoG without causal nodes. Yang et al. learn concrete action commands from small video clips. Unlike their system, our design allows a modifiable grammar and our performance is measured on multi-step actions.

Contributions

The contributions of our paper include the following:

- A unified real-time framework for learning and inference on a robot system by using an STC-AoG data structure.
- A formal definition of knowledge transfer with respect to human-computer interactions.
- A test to evaluate the success of a human-robot knowledge transfer system.

Our Approach

We encapsulate knowledge by an expressive graphical data structure: Spatial-Temporal-Causal And-Or Graph (STC-AoG) $G_\Omega = (G_s, G_t, G_c)$ which models the compositional structures of objects, actions and tasks. A specific piece of information or skill, such as how to fold clothes, is a subgraph $G \subseteq G_\Omega$. The goal of knowledge transfer is to deliver G from one agent (e.g. human) to another (e.g. robot) with a minimum loss of knowledge.

In human-robot interactions, we restrict communication to only physical actions through a video-camera sensor V, and natural language text L. Therefore, the learner must construct an optimal G based only on V and L, resulting in the Bayesian formulation,

$$G^* = \arg\max_{G_t} P(G_t|V, L) = \arg\max_{G_t} \frac{P(V|G_t, L)P(G_t, L)}{P(V, L)}$$

Similar to (Ha, Kim, and Zhang 2015), we use a graph Monte Carlo method that assumes the graph structure is determined only by that of the previous iteration.

$$G^* = \arg\max_{G_t} P(V|G_t, L)P(G_{t-1}, L)$$

The learning algorithm is similar to a marginalization-based parameter learning algorithm, where we first marginalize our STC-AOG, and learn the S-AOG, T-AOG and C-AOG separately, then jointly learn the conditional model between each other.

Figure 2 shows a small segment of G^*, and specific details of the spatial, temporal, and causal segments are described as follows.

Spatial Representation

Our system encodes sensory data from the environment to form a belief representation. We use a Kinect camera to capture both the color image and point cloud per frame. For our cloth folding task, we built a cloth detector and defined a compact encoding. Specifically, we represent every cloth by a high-level understanding based off the bounding and inscribed rectangles of the cloth contour, and a low-level understanding based off specific keypoints. The keypoints

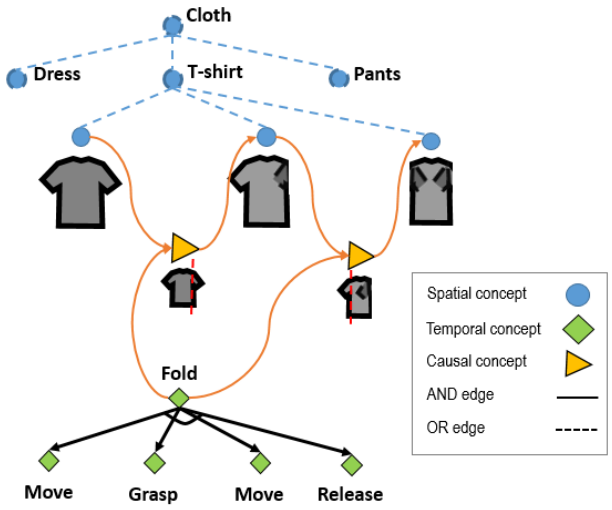

Figure 2: An automatically learned STC-AoG.

and contour shape data are used as input to the folding algorithm which generalizes to arbitrary articles of clothing. To store the hierarchical structure of physical objects, we use an And-Or Graph data-structure, called the Spatial And-Or Graph (S-AoG) (Zhu and Mumford 2006). **AND** nodes in the S-AoG represent structural compositionality (i.e. a vehicle has an engine). **OR** nodes in the S-AoG represent variations (i.e. a car is a type of vehicle).

Causal Representation

The perceived model of the world is then used to learn a logical cause-and-effect type of reasoning from a single-instance, inspired by the Dynamics Model (Wolff 2007).

The Dynamics Model defines causal relationships as interpretations of force vectors. The nodes in the S-AoG are normalized feature vectors in a higher dimensional space, and are acted on by force vectors from the T-AoG. As per the model, if the net force on a spatial node is collinear with the vector represented by the end-state of an action, then a causality is deduced, as shown in Figure 3.

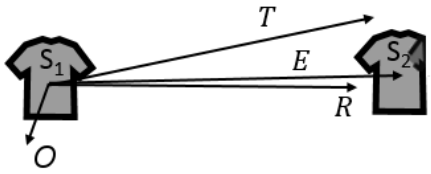

Figure 3: Forces associated with the simulated affector T and other forces O sum to produce a resulting R force, which is nearly collinear to the actual end-state E, implying that T causes S_1 to become S_2.

The causal relationships are stored in a Causal And-Or Graph (C-AoG). **AND** nodes in the C-AoG indicate that all

preconditions are necessary, whereas **OR** nodes indicate that only one of the preconditions is sufficient.

Temporal Representation

These deductive models are used to plan out the next course of action, which may affect the environment. The actuators that affect the environment, whether by the robot or the human, are represented in another data-structure, called the Temporal And-Or Graph (T-AoG). **AND** nodes represent actions done in a series of steps. **OR** nodes represent variations in possible actions.

Joint Representation

We represent the belief models (S-AoG), the reasoning models (C-AoG), and the environment actuators (T-AoG) all into one unified Spatial Temporal Causal And-Or Graph (STC-AoG) data structure. As a consequence, the whole system forms a closed-loop from perception to learning to inference, and back again to perception. Figure 2 demonstrates a small portion of the STC-AoG applied to a cloth-folding task.

Knowledge Transfer Test

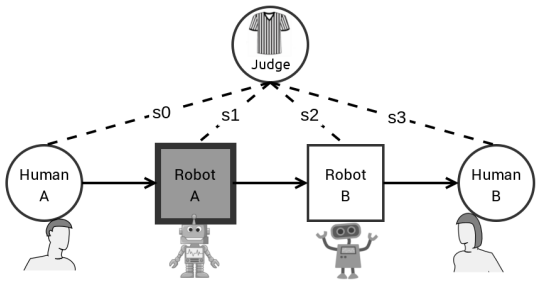

Figure 4: Arrows represent the direction of knowledge transfer. The judge assigns task scores at each step.

To determine the proficiency of knowledge transfer to and from an artificial agent, we propose a three-part test.

A human demonstrator H_A will perform a chosen task to receive a task score s_0 by a human judge. In the first part of the test, H_A will teach this task to a robot R_A that has not been previously trained on the task. The judge will assign a task score s_1 based on R_A's performance.

Next, the second test will evaluate R_A's ability to transfer knowledge to another robot R_B that has not been previously trained on the task. Robot-to-robot knowledge transfer can be as direct as sending over the explicit knowledge structure, which in our case is the STC-AoG. Again, the judge will assign a task score s_2.

Finally, R_B must teach the task to a human H_B that has no previous knowledge of the task procedure. A task score s_3 will be assigned by the judge. If all three task scores match within 10% of s_0, then R_A is said to have passed the Knowledge Transfer Test (KTT). The entire process is visualized in Figure 4

Experimental Results

We evaluate our framework on a two-armed robot using the proposed Knowledge Transfer Test on a cloth folding task. For simplicity, the robot learns using only video input, in which a human demonstrates how to fold a t-shirt.

To benchmark real-time performance, we calculate the ratio between the duration of the demonstration and the total time spent learning. The average speed of our robot system is 10 fps, resulting in a system which is nearly real-time, outperforming most perception-heavy robot learning-systems.

Our robot was able to understand the cloth-folding task, generating a STC-AoG similar to Figure 2, confidently enough to pass the first part of the KTT. We were able to save the graphical structure and load it into a fresh robot to pass the second part of the KTT. The robot was also able to teach the task successfully to a human, but since folding clothes is already a well known skill by most humans, we set aside deeper investigation of robot-to-human teaching for future work.

Acknowledgments

This work was supported by DARPA's Simplifying Complexity in Scientific Discovery (SIMPLEX) grant. In addition, we would like to thank SRI International and OSRF for their hardware support.

References

Blum, A. L., and Furst, M. L. 1997. Fast planning through planning graph analysis. *Artificial Intelligence* 90(1):281–300.

Gupta, A.; Srinivasan, P.; Shi, J.; and Davis, L. S. 2009. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *CVPR*, 2012–2019. IEEE Computer Society.

Ha, J.-W.; Kim, K.-M.; and Zhang, B.-T. 2015. Automated construction of visual-linguistic knowledge via concept learning from cartoon videos. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015), Austin*.

Miller, S.; van den Berg, J.; Fritz, M.; Darrell, T.; Goldberg, K.; and Abbeel, P. 2012. A geometric approach to robotic laundry folding. *International Journal of Robotics Research (IJRR)* 31(2):249–267.

Si, Z.; Pei, M.; Yao, B.; and Zhu, S.-C. 2011. Unsupervised learning of event and-or grammar and semantics from video. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 41–48. IEEE.

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.

Wolff, P. 2007. Representing causation. *Journal of experimental psychology: General* 136(1):82.

Xiao, T.; Zhang, J.; Yang, K.; Peng, Y.; and Zhang, Z. 2014. Error-driven incremental learning in deep convolutional neural network for large-scale image classification. In *Proceedings of the ACM International Conference on Multimedia*, 177–186. ACM.

Yang, Y.; Li, Y.; Fermuller, C.; and Aloimonos, Y. 2015. Robot learning manipulation action plans by unconstrained videos from the world wide web. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

Zhu, S.-C., and Mumford, D. 2006. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision* 2(4):259–362.