Interactive Robot Knowledge Patching using Augmented Reality

Hangxin Liu\textsuperscript{1}\textsuperscript{*}, Yaofang Zhang\textsuperscript{1}\textsuperscript{*}, Wenwen Si\textsuperscript{2}, Xu Xie\textsuperscript{1}, Yixin Zhu\textsuperscript{1}, and Song-Chun Zhu\textsuperscript{1}

Abstract—We present a novel Augmented Reality (AR) approach, through Microsoft HoloLens, to address the challenging problems of diagnosing, teaching, and patching interpretable knowledge of a robot. A Temporal And-Or graph (T-AOG) of opening bottles is learned from human demonstration and programmed to the robot. This representation yields a hierarchical structure that captures the compositional nature of the given task, which is highly interpretable for the users. By visualizing the knowledge structure represented by a T-AOG and the decision making process by parsing the T-AOG, the user can intuitively understand what the robot knows, supervise the robot’s action planner, and monitor visually latent robot states (e.g., the force exerted during interactions). Given a new task, through such comprehensive visualizations of robot’s inner functioning, users can quickly identify the reasons of failures, interactively teach the robot with a new action, and patch it to the current knowledge structure. In this way, the robot is capable of solving similar but new tasks only through minor modifications provided by the users interactively. This process demonstrates the interpretability of our knowledge representation and the effectiveness of the AR interface.

I. INTRODUCTION

The ever-growing vast amount of data and rapid-increasing computing power have enabled a data-driven machine learning paradigm in the past decade. Using Deep Neural Networks (DNNs) \cite{1}, the performance of machine learning methods has reached a remarkable level in some specific tasks, even arguably better than human, \textit{e.g.}, control \cite{2}, \cite{3}, grasping \cite{4}, \cite{5}, object recognition \cite{6}, \cite{7}, learning from demonstration \cite{8}, and playing the game of go \cite{9} and poker \cite{10}, \cite{11}. However, despite these recent encouraging progress, DNN-based methods have well-known limitations; one of these limitations is the lack of interpretability of the knowledge representation, especially about how and why a decision is made, which plays a vital role in the scenarios where robots work alongside humans.

Meanwhile, contextual adaptation models using And-Or-Graph (AOG) and Probabilistic Programming start to demonstrate the interpretability using small amount of training data in robot learning \cite{12}, \cite{13}, recognition \cite{14}, \cite{15}, reconstruction \cite{16}, social interactions \cite{17}, causal reasoning \cite{18}, \cite{19}, playing video games \cite{20}, and human-level concept learning \cite{21}. Although these types of models have been identified by DARPA as the representative models in the third wave of artificial intelligence \cite{22}, a natural and convenient way to teach and interact with a robot to acquire and accumulate such interpretable knowledge is still missing.

In this paper, we propose an augmented reality (AR) interface, through Microsoft HoloLens, to interact with a Re-think Baxter robot for teaching and patching its interpretable knowledge represented by the AOGs. In the experiments, we demonstrate the proposed AR interface develops interpretations at three different levels:

1) Knowledge structure by compositional models. We take an example of a robot opening various medicine bottles, and represent the robot’s knowledge structure using a Temporal And-Or Graph (T-AOG) \cite{14}. The T-AOG encodes a repertoire of a successful action sequence for a robot to open medicine bottles. Visualizing through the holographical interface, the state of robot represented by a T-AOG can be naturally inquired through gesture control (see Fig. 7a).

2) Interpretable decision making. Unlike a teacher can usually query students to verify whether they obtain the knowledge structure correctly, it is nontrivial for users to check and understand robots’ inner functioning, making it difficult for users to diagnose the robot decision-making process. By visualizing the decision-making process on top of T-AOG through the holographical interface, information of interests can be better associate to the actual

\textsuperscript{*} Hangxin Liu, Yaofang Zhang contributed equally to this work.

\textsuperscript{1} Hangxin Liu, Yaofang Zhang, Xu Xie, Yixin Zhu, and Song-Chun Zhu are with UCLA Center for Vision, Cognition, Learning, and Autonomy at Statistics Department. Emails: \{hx.liu, v.zhang, xiexu, yixin.zhu\}@ucla.edu, sczhu@stat.ucla.edu.

\textsuperscript{2} Wenwen Si is with Department of Automation, Tsinghua University. Email: sww12@mails.tsinghua.edu.cn.
robot and the actual scene, thus help to gain insight about how a robot behaves and why it behaves in a certain way.

3) Interactive knowledge structure patching. Once the users find out the reason why a certain action sequence leads to a failure, the users can interactively patch the knowledge structure represented by the T-AOG: adding a missing node, deleting a redundant node, and updating a node representing a wrong action.

A. Related Work

Explanation Interfaces has a long and rich history in artificial intelligence. The goal of explanation interfaces is to generate explanations regarding particular predictions and decisions automatically so that users can diagnose and correct the wrong behaviors. Such types of systems have been deployed in a wide range of applications, e.g., medical diagnoses [23], understanding agent’s action [24], [25], activity recognition [26], robot control [27], and simulator for training in virtual environment [28], [29]. By adopting HoloLens, the most advanced commercial AR product to date, we hope to provide a fully mobile yet powerful explanation interface with modern visualizations, faster diagnoses, and more natural interactions.

Augmented Reality (AR) can overlay the symbolic and semantic information of a robot either to a simulator [30], [31], [32] or to a real-world scene. In particular, [33], [34], [35], [36] conveyed robots intention (e.g., movements and trajectory) through projecting visual aids. Krückel et al. [37] transferred robot’s camera view to user’s head-mounted display in order to achieve better teleoperation. In other cases, researchers deploy AR to display robot’s states to help users gain insight of the multi-robot systems [38], [39]. Compared to the present study, this line of work requires a fixed camera and/or projector, limiting the mobility of both human and robots.

AR is also proven to be effective in interacting with robots. Brageul et al. [40] incorporated AR techniques to develop a user interface to program a robot. Zaeh et al. [41] used a laser pointer with AR to update robot’s trajectory. Kuriya et al. [42] introduced infrared marker projection and detection pipeline for robot navigation. Huy et al. [43] proposed a comprehensive system that consists of laser-writer, see-through head-mounted display, and a hand-held device to control a robot. Anderson et al. [44] enabled a robot to project its task information, e.g., welding points, to the task space for user verification and re-programming. Compared to the present study, prior efforts mainly focused on providing one-time guidance, but not for teaching and accumulating interpretable knowledge for future similar tasks.

Interpretability of a robot’s planning and knowledge representation determines whether users can effectively understand, verify, diagnose and agree with robot’s behaviors. Past work in robot’s interpretable representation can be summarized as two types. The first kind of representation utilizes Markov Decision Processes (MDP): it depicts the state transitioning by performing an action associated with a reward. For instance, Feng et al. [45] helped operator reason about system violations by defining a notion of structured probabilistic counterexamples. Hayes and Shah [46] used MDP to generate a verbal explanation of robot behaviors. However, the decision rules obtained by the MDP is not readily interpretable as it lacks long-term dependencies.

The second type is the graphical models, including Hierarchical Task Network (HTN) that is widely used in robotics, and And-OR Graph (AOG) rooted from the community of knowledge representation. This type of models symbolically abstracts each motion primitive. It has been utilized in cloth-folding with causal relation [12], human-robot collaboration [47], learning social interactions [17], and complex manipulation for opening medicine bottles [48].

Learning from Demonstration (LfD) is a vast field with a rich history [8]. In the present study, using vision-based algorithms, the robot can learn the action sequence of opening bottles from the demonstrations. Given demonstrations of opening a medicine bottle with safety lock, the additional action “push” is not directly perceivable from human demonstrations using vision sensor alone, making the demonstrations identical to the ones for conventional bottles, resulting in the failure of opening the lid. In such cases, the users can use the proposed AR interface to interactively diagnose the action sequence and correct the missing action, leading to a successful opening.

B. Contribution

This paper makes the following three contributions:

1) We introduce a new AR interface based on the state-of-the-art head-mounted display, Microsoft HoloLens, providing users a much more natural way to interact with a robot. In addition to visualizing robot’s states, intentions, or controlling robots, we further visualize robot’s knowledge representation so that users can understand why and how a robot will behave.

2) In contrast to using additional force sensing [49] to perceive the visually hidden force [48], the present study provides an intuitive way for users to augment the visually imperceptible knowledge on top of the learned action sequence represented by a T-AOG. In this way, the AR interface affords a much more effectively diagnose and knowledge patching process. Furthermore, it often has a much lower cost, as users do not need to build any additional sensors or apparatus to demonstrate the tasks.

3) We build a communication interface between the HoloLens platform and ROS, and are publicly available online. It allows a variety of interchangeable messages, which we hope would ease the development difficulties across commonly used platforms.

C. Overview

The remainder of the paper is organized as follows. Section II outlines the AOG representation and the architecture of AR system. Learning knowledge representation of opening bottles from human demonstration is described in Section III. In Section IV, we showcase some experiment results, including the visualization of the interpretable knowledge, the

https://github.com/xiaozhuchacha/AOG_AR
Fig. 2: Illustration of a T-AOG. The T-AOG is a temporal grammar in which the terminal nodes are motion primitives of hand-object interactions.

**II. METHODOLOGY**

This section introduces the representation of knowledge structure and describes the proposed AR system architecture.

**A. Representation**

We represent the action sequence to execute a task by a structural grammar model, T-AOG (see Fig. 2). An AOG is a directed graph which describes a stochastic context-free grammar (SCFG), providing a hierarchical and compositional representation for entities. Formally, an AOG is defined as a five-tuple $G = (S, V, R, P, \Sigma)$. Specifically,

- $S$ is a start symbol that represents an event category (e.g., opening a bottle).
- $V$ is a set of nodes including the non-terminal nodes $V_{NT}$ and terminal nodes $V_T$: $V = V_{NT} \cup V_T$. The non-terminal nodes can be divided into And nodes and Or nodes: $V_{NT} = V_{And} \cup V_{Or}$. An **And-node** represents the compositional relations: a node $v$ is an And-node if the entity represented by $v$ can be decomposed into multiple parts, which are represented by its child nodes. An **Or-node** indicates the alternative configuration among its child nodes: a node $v$ is an Or-node if the entity represented by $v$ has multiple mutually exclusive configurations represented by its child nodes. The **terminal** nodes $V_T$ are the entities that cannot be further decomposed or have different configurations; it represents the set of motion primitives that a human/robot can perform (e.g., approaching, twisting).
- $R = \{r: \alpha \rightarrow \beta\}$ is a set of production rules that represent the top-down sampling process from a parent node $\alpha$ to its child nodes $\beta$.
- $P : p(r) = p(\beta | \alpha)$ is the probability associated with each production rule.
- $\Sigma$ is the language defined by the grammar, i.e., the set of all valid sentences that can be generated by the grammar.

A **parse tree** is an instance of AOG, where for each Or-node, one of the child nodes is selected. A temporal parse tree $pt$ of an event is a sub-graph of the T-AOG that captures the temporal structure of the scenario. The terminal nodes of $pt$ form a valid sentence; in this case, terminal nodes are a set of atomic actions for the robot to execute in a fixed order.

**B. System Architecture**

**AR Headset:** Using AR headset, both symbolic and semantic information of the robot can be augmented on top of the observed actual scenes, allowing users to gain better situational awareness and insights of the robot’s status. In the present work, we adopt the state-of-the-art AR head-mount display HoloLens. Compared to other available AR headsets, HoloLens is the first untethered AR head-mounted display that allows the user to move freely in the space without being constrained by any cable connections. Integrated with 32-bit Intel Atom processors, HoloLens provides a reliable localization using IMU, four spatial-mapping cameras, and a depth camera. Using Microsoft’s Holographic Processing Unit, the users can realistically view the augmented contents. Common interactions, such as gaze, hand gesture, and voice control with Cortana, are integrated, making HoloLens the most suitable device for the present study. The holograms displayed on its screen are created using Unity3D game engine, through which various visual effects can be introduced.

**Robot Platform:** We experiment the proposed AR interface with a robot platform consists of a dual-armed 7-DoF Baxter robot from Rethink Robotics mounted on a Data Speed mobility base. The robot is equipped with a ReFlex TackkTile gripper on the right wrist and a Robotiq S85 parallel gripper on the left. The entire system runs on Robot Operating System (ROS), and arm motion planning is computed using MoveIt! This comprehensive research robotics system has been proven suitable for many challenging tasks in robotics researches [12], [17], [48].

**Overall Framework:** Fig. 1 illustrates the system architecture of the proposed interface. A repertoire of robot opening a conventional bottle (Bottle 1) with no safety lock is taught to the robot through imitation learning. We visualize its interpretable knowledge representation and a set of AR elements that reveal the robot’s state and decision-making process represented by a parse tree $pt$ through the HoloLens.

Given a new scenario, i.e., to open several medicine bottles (Bottle 2-4) that require pressing down the lid during twisting, the robot is asked to execute and open these bottles. For such task, the critical actions involved (e.g., whether pressing down or not) are perceptually similar to the actions of opening conventional bottles. Therefore, the learned knowledge of opening a conventional bottle becomes insufficient; without knowing the action of pressing down, the sampled $pt$ always leads to failures of opening medicine bottles with safety lock.

Through the AR interface, users can quickly identify the reasons for failures through HoloLens, interactively teach the robot with a new action, and patch it to the knowledge...
structure represented by the T-AOG. In this way, the robot is capable of solving similar but new tasks only through minor modifications provided by the users interactively. This process demonstrates the interpretability of our knowledge representation and the effectiveness of the AR interface.

III. Imitation Learning

This section briefly describes the pipeline for teaching a robot to open conventional bottles through imitation learning. We adopt the pipeline proposed by Edmonds et al. [48]. In contrast, our work makes two differences: i) we capture pose information of hand-object interactions with LeapMotion sensor to avoid the use of a tactile glove [49] and Vicon, ii) we apply a modified version of the ADIOS (automatic distillation of structure) [50] to induce the temporal grammar T-AOG of the task, yielding a more compact AOG model.

Human Data Collection: Twenty manipulations of opening medicine bottles are collected using hand tracking by LeapMotion sensor. The captured data is manually segmented into sequences of atomic actions. Fig. 3 presents the average hand skeleton of each atomic action, as well as the description of these atomic actions. Note that two actions are different during contacting the lid: Grasp refers to a grasping with all fingertips contacting with the lid, whereas Pinch is the action that only the tips of the thumb and the index finger contact with the lid.

Grammar Induction: A T-AOG is induced from segmented action sequences using a modified version of ADIOS algorithm [50]. It results in a stochastic context-free grammar with probabilistic Or-nodes (see Fig. 4). Given the learned T-AOG, a parse tree \( pt = (a_0, \ldots, a_K) \) can be obtained by decomposing all the And-nodes and selecting one branch at each Or-node. The robot can execute \( pt \) by performing the actions encoded by the terminal nodes in temporal order to accomplish the task.

Mirroring Human Actions to Robot: We endow the robot with a dictionary of atomic actions corresponding to the human’s manipulative actions (see Fig. 3). Specifically, each action is represented by the change of robot’s end-effector pose or the open/close of the gripper. For instance, the robot approaches the lid by moving to a new pose assuming the relative pose of the lid is known, twists the lid by rotating the gripper to the counter-clockwise direction, and moves the lid to rotate in the opposite direction.

IV. State Visualization and Knowledge Patching

This section showcases the functionality and the effects using the proposed AR interface: i) diagnosing the formerly obscure robot inner functioning and knowledge structure becomes possible through the visualization, and ii) the user can teach new knowledge to the robot in novel scenarios by patching the interpretable knowledge structure.

Notice that the following qualitative results are captured in two ways. The first type of results is captured by a DSLR camera seeing through the HoloLens. This is what a user would see directly through the HoloLens, and it mimics the first-person egocentric view with distortion, slightly worse color contrast, and some blurs (Fig. 5a, 5b, 5d, and 6a) due to limited field of view, the curvature of the lens, and reflectance of HoloLens’ screen. The second type of results are captured by the mixed reality capture feature of the HoloLens that overlays the holograms to the image captured by HoloLens’s PV camera (Fig. 5c, 5e, 6b, and 6c). Although the second type of the images is of higher quality, they are not exactly the same as what users would see. Hence, we present the results in Fig. 5c and 6c in a mixture format using two capturing methods to show the actual realism that the HoloLens would afford to users in real-world, whereas Fig. 7 to 10 use images captured only by the second type to better illustrate the functionality and the procedures of the proposed AR interface.

A. AR Interface

We developed a two-way communication bridge between ROS and HoloLens to exchange various messages across two platforms. From ROS to HoloLens, the bridge allows the ROS topic messages generated by the robot to be
By default, if the sensory information is turned on, the panel showing the information will be displayed next to the actual sensor in 3D scene (see Fig. 6a and 6b). The locations of these panels can be freely dragged to any 3D positions around the robot using gesture control (see Fig. 6d). Compared to other AR applications which display information through tables or cell phones in which users’ hands are occupied, the present AR interface frees the users’ hands, providing better interactive experience as well as affording more complex and natural gesture controls.

B. Visualizing Knowledge Structure and Decision Making

In addition to augmenting AR elements shown in Fig. 5 and 6 we also reveal the robot’s inner functioning and knowledge structure through the holographical interface. The knowledge structure is represented by a T-AOG (see Fig. 7a), which encodes a repertoire of opening bottles. The structure of the T-AOG can be naturally inquired through gesture control. This feature provides users with a high-level semantic understanding of the robot’s action planner, that is, how the robot behaved and will behave later.

Parsing the T-AOG will produce an action sequence (see Fig. 7b), which consists of atomic actions that the robot can execute to fulfill the task. By closely monitoring the dynamic parsing process, the users can supervise the decision-making process of the robot. As an example, Fig. 7d and 7e demonstrate three representative steps in the parsing. Next action is selected with 100% at an And-node (see Fig. 7d) as it represents a compositional relation and its child nodes are deterministically executed in a temporal order. An Or-node (see Fig. 7c and 7e) indicates a switching configuration among its child nodes; one of its child nodes is selected based on the branching probability.

By visualizing the sensory data (see Fig. 5, 6, and 8), together with the knowledge structure and decision making process (see Fig. 7), the robot’s inner functioning is revealed comprehensively.

C. Diagnoses, Motion Control, and Knowledge Patching

Using the features provided by the proposed AR interface, users can understand why (T-AOG parsing) and how (sensory information, e.g., force response) the robot behaves. An example is provided in Fig. 10a showing an execution process in opening a conventional bottle Bottle 1. Additionally, users can diagnose redundant or wrong behaviors during the executions.
Fig. 7: Visualizing the robot’s knowledge structure and decision making process to the users. (a) The knowledge representation, T-AOG. (b) A valid action sequences for opening bottles generated from the T-AOG. The robot decides the next action by parsing the T-AOG: (c) (e) the next action is select by the branching probability at Or-node, and (d) the next action is planned deterministically at And-node.

Fig. 8: Visualizing the force exerted by the left end-effector shows the force readings of different actions: (a) grasp the Bottle 1, (b) pinch the Bottle 1, (c) twist the Bottle 1, (d) grasp the Bottle 2, (e) twist the Bottle 2, and (f) push the Bottle 2.

Fig. 9: User programs a new action through AR interface by dragging the virtual gripper model to a new pose. (a) A warning sign appears if the new pose is too aggressive. (b) A new action Push is successfully programmed with proper pose and parameters.

Fig. 8 demonstrates a typical diagnose process in which a user may reason about the redundant behaviors. Specifically, Fig. 8a, 8b, and 8c show the robot’s force readings of its left end-effector when performing the grasp, pinch, and twist action in opening conventional bottle Bottle 1, respectively. The red, green, and blue arrows indicate the canonical x, y, z-direction relatively to the robot’s base, and the length of these arrows are proportional to the corresponding force magnitude sensed by the force/torque sensor at the end-effector. The white arrow is the vector sum of these forces, providing a more intuitive indication of how the force is applied. Not only are the grasp and pinch actions executed in the same way (closing the fingers), but also the force responses are identical. These readings come to the conclusion that the switching configuration between grasp and pinch is redundant and one of the actions could be removed.

Wrong actions can also be discovered. When a new task is given, i.e., opening a medicine bottle with child-safety lock that requires an additional pressing-down action on the lid (e.g., Bottle 2-4), the robot executes the action sequence based on existing knowledge (see Fig. 10b). Although the majority of the robot’s existing knowledge remains unchanged and produces desired results (see Fig. 8d), it lacks the concept of “pressing down action”, resulting in the failure to complete the task: the twist action only applies small downward force, trying to unlock the safety mechanism.

In contrast to the traditional methods that require huge efforts in re-programming or re-training the robot for new repertoire, the proposed AR interface allows users to easily provide new guidance interactively without physically interacting with the robot. This teaching process is accomplished by dragging the virtual gripper hologram to a new pose (see Fig. 9). A warning sign would show up to alert users if the new pose is out of range (see Fig. 9a). By locking the displacements in certain directions, the user can define a new action through modifying the existing twist action by moving the end-effector downward to produce pressing force, namely push action (see Fig. 9b). The resulting force reading is shown in Fig. 8f. This process not only simplifies the teaching process but also avoids physical contact between users and the robot, which is safer to both human and robots.

Instead of providing single-time guidance to the robot, we can interactively modify its knowledge so that the acquired skill can be stored for future similar tasks. Due to the merit of T-AOG’s high interpretability, the knowledge patching is accomplished by deleting a redundant node or adding a missing node. Fig. 10c shows an execution after patching the knowledge, which leads to a successful execution. The
A number of ROS topic messages across different robot systems, such as TF, and images can be visualized to users. By visualizing the robot’s knowledge structure represented by a T-AOG and its decision-making process, users can gain a better insight of its inner functioning. When a new task is given, the proposed AR interface allows easy diagnosing about the failure execution, as well as providing an interactive approach to teach the robot with new knowledge, as well as maintain such knowledge by patching the T-AOG. Overall, the proposed interface boosts the interactions between the users and the robot.

There are several aspects of the future work that can substantially improve the proposed AR interface. As the demonstration data is current collected using a vision-based method, learning the task composition (T-AOG) can be achieved by unsupervised learning such that the robot can acquire most of the knowledge by simply observing demonstrations without any manual data labeling. In addition to gesture interactions through HoloLens, language interactions can be explored as well to provide more effective and natural interactions. Lifelong learning and active learning are promising as it allows robots ask user questions in their decision-making process so that the knowledge patching can take place autonomously by querying, which is closer to humans teaching-learning framework. Additionally, it is also interesting to collect human ratings based on the interpretability of various methods.

**Acknowledgement:** The authors thank Zhenliang Zhang and Feng Gao for their assistance in the experiment. The work reported herein is supported by DARPA XAI N66001-17-2-4029 and ONR MURI N00014-16-1-2007.