2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
September 24–28, 2017, Vancouver, BC, Canada

Feeling the Force: Integrating Force and Pose for Fluent Discovery through Imitation Learning to Open Medicine Bottles

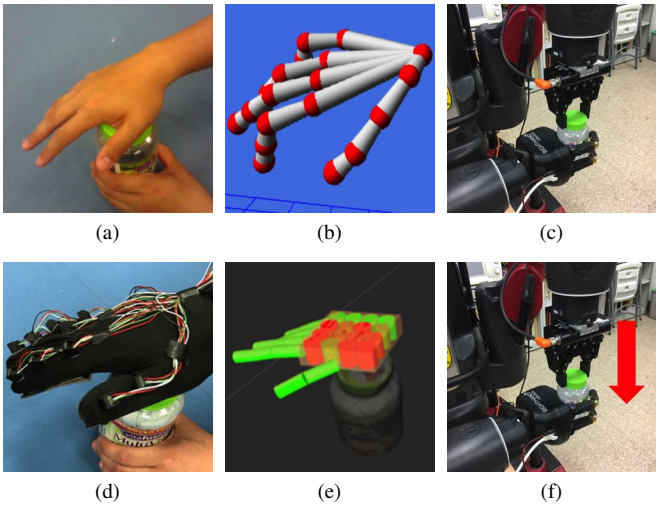
Mark Edmonds[1⋆], Feng Gao[1⋆], Xu Xie[1], Hangxin Liu[1]
Siyuan Qi[1], Yixin Zhu[1], Brandon Rothrock[2], Song-Chun Zhu[1]

***Abstract*— Learning complex robot manipulation policies for real-world objects is challenging, often requiring significant tuning within controlled environments. In this paper, we learn a manipulation model to execute tasks with multiple stages and variable structure, which typically are not suitable for most robot manipulation approaches. The model is learned from human demonstration using a tactile glove that measures both hand pose and contact forces. The tactile glove enables observation of visually latent changes in the scene, specifically the forces imposed to unlock the child-safety mechanisms of medicine bottles. From these observations, we learn an action planner through both a top-down stochastic grammar model (And-Or graph) to represent the compositional nature of the task sequence and a bottom-up discriminative model from the observed poses and forces. These two terms are combined during planning to select the next optimal action. We present a method for transferring this human-specific knowledge onto a robot platform and demonstrate that the robot can perform successful manipulations of unseen objects with similar task structure.**

I. Introduction

Consider the task of opening medicine bottles that have child-safety locking mechanisms (Fig. 1(a)). These bottles require the user to push or squeeze in various places to unlock the cap. By design, attempts to open these bottles using a standard procedure will result in failure. Even if the agent visually observes a successful demonstration, imitation of this procedure will likely omit critical steps in the procedure. The visual procedure for opening both medicine and traditional bottles are typically identical. The agent lacks understanding of the tactile interaction required to unlock the safety mechanism of the bottle. Only direct observation of forces or instruction can elucidate the correct procedure (Fig. 1(e)). Even with knowledge of the correct procedure, opening medicine bottles poses several manipulation challenges that involve feeling and reacting to the internal mechanisms of the bottle cap. Although the presented study takes opening medicine bottles as an example, many other tasks share similar properties and require non-trivial reasoning such as opening locked doors [1].

In this paper, we learn a manipulation model from human demonstration that captures observed motion and kinematics, as well as visually latent changes such as forces and internal state (Fig. 1(e)). We learn this manipulation model for objects that have similar functional properties, but exhibit different geometries and internal configurations that affect how the object must be manipulated.



(a) (b) (c) (d) (e) (f)

Fig. 1: Given a RGB-D-based image sequence (a), although we can infer the skeleton of hand using vision-based methods (b), such knowledge cannot be easily transferred to a robot to open a medicine bottle (c), due to the lack of force sensing during human demonstrations. In this work, we utilize a tactile glove (d) and reconstruct both forces and poses from human demonstrations (e), enabling robot to directly observe forces used in demonstrations so that the robot can successfully open a medicine bottle (f).

Two key problems are discussed in this paper:

1) how to naturally recover the visually latent force data from the human demonstrations, and
2) how to represent such knowledge and successfully transfer it to a robot?

For the first problem, although some initial results have been reported to reconstruct poses and/or forces exerted by the demonstrator using vision-based methods [2], [3], [4], [5], [6], these methods still have difficulty providing pose and force data precise enough for robot learning. Instead, we utilize an open-source tactile glove [7] designed to measure both hand pose and contact forces across the surface of the hand. These demonstrations are performed naturally, and within a motion capture setup to obtain ground-truth tracking of the objects and human wrist.

For the second problem, our system takes into consideration: i) an And-Or-Graph (AOG) [8] learned from human demonstrations as top-down knowledge for manipulations of an unseen medicine bottle, in which the AOG model uses *fluents* [9] to model the changes between pre- and post-

⋆ Mark Edmonds and Feng Gao contributed equally to this work.

[1] UCLA Center for Vision, Cognition, Learning, and Autonomy at Statistics Department. Emails: {markedmonds, f.gao, xiexu, hx.liu, syqi, yixin.zhu}@ucla.edu, sczhu@stat.ucla.edu.

[2] Jet Propulsion Laboratory, California Institute of Technology. Email: rothrock@jpl.nasa.gov.

conditions of demonstrations in a low-dimensional subspace; and ii) A bottom-up process learned from raw signal data when robot executes to encode transition between pre- and post-conditions. Together, these two processes learn a manipulation model to open medicine bottles.

A. Related Work

Tactile Gloves are common tools to capture demonstration data [10]. In this paper, we use a tactile glove [7] to record both human pose and visually hidden forces applied at each proximal and distal phalange, as well as a 4-by-4 grid of sensors to detect forces exerted by the palm. In the literature, most data gloves use IMUs [11], [12], [13] or curvature sensors [14], [15] to track finger pose. To read force, FlexiForce [16] sensors or Velostat [17], [18], [19] are commonly adopted.

Learning from demonstration (LfD) is a crucial component to building general purpose robots, and a very broad field with rich history. This literature is too expansive to survey here; we refer readers to a survey [20]. Instead, we focus on approaches related to our work: kinesthetic teaching, teleoperation, and imitation learning in the next paragraphs. Note that humans are able to learn quickly from one or only a few examples for a new task [21], thus teaching robots to achieve similar performance would enable robots to enter many routine human activities. In this paper, our approach requires a relatively small number of examples, approximately 10 examples per bottle.

Kinesthetic teaching and teleoperation both enable direct mappings between demonstrations and executions [20] and have successfully demonstrated capability of learning both motor skills [22], [23] and manipulation policies [24], [25]. However, this direct embodiment mapping, a typically complex function that maps states/actions in demonstrations to states/actions on the robot [20], is ill-suited for manipulation tasks that incorporate forces. Although some robots have built-in force sensing, the demonstrator often cannot receive feedback from forces applied. To address this problem, Kormushev *et al.* [26] used kinesthetic teaching to demonstrate positional requirements of a task and employed a secondary haptic demonstration to provide required forces. In contrast, our approach simultaneously integrates both poses and forces within a single demonstration using a tactile data glove, providing a more natural and efficient way to sense force from a demonstration.

Imitation learning has two main streams: i) behavior cloning through supervised demonstrations that directly mimic the demonstrator's behaviors [27], [28], [29], [30], [31], [32], [33], [34], and ii) inverse reinforcement learning [35], [36], [37]. While inverse reinforcement learning is limited to Markvoian problems, our approach falls into behavior cloning and is capable of handling both Markovian and non-Markovian problems by utilizing a grammar model.

Two previous work stands out as most relevant to the presented work. Huang *et al.* [38] use imitation learning coupled with a data glove for opening a set of standard bottles without understanding the internal configuration. This simplification is infeasible when dealing with locking mechanisms of medicine bottles, which require direct and complex manipulation of the cap beyond pure rotation. Sung *et al.* [39] uncovered haptic components of a task from teleoperated demonstrations. In contrast, our work learns the manipulation tasks directly from human demonstration using a tactile glove, resulting in more natural and larger variety of demonstrations. In addition, Sung *et al.* used a recurrent neural network based method that typically only encodes a few steps of dependencies. However, our work uses an explicit grammar to generate actions, capable of incorporating long-term temporal dependencies.

B. Contribution

This paper makes four contributions:

1) Using a tactile glove during demonstrations that enable the robot to utilize both the poses and forces exerted by the demonstrator. In contrast with previous work, our method focuses on integrating visual measurements with physical measurements not observable from vision (*e.g.* forces), capturing latent relationships that are imperceptible from vision alone.
2) Learning a stochastic grammar model that represents the compositional task hierarchy comprising of atomic actions for manipulation tasks, compactly capturing the admissible sequence of actions for all the bottles demonstrated.
3) Learning a bottom-up process that encodes raw haptic signals to account for the transition from a previous state to a new state. Together with the stochastic grammar model as a top-down process, these two processes jointly form the manipulation model.
4) Transferring the learned model from human demonstrations onto a Baxter robot by solving a correspondence problem [40]. This embodiment mapping function directly relates hand pose and contact force from the human to the force-torque sensing and gripper state of the robot; enabling the robot to reason about its haptic measurements using the relations learned from human demonstration.

C. Overview

In this paper, we use human demonstrations to learn a manipulation model based on an AOG representation that integrates both poses and forces. Section II outlines the AOG representation and related components. Section III discusses our data collection environment, instruments, and procedures. In Section IV, we present how to learn an AOG representation from demonstrations, and how to combine it with raw signals using a bottle-up process to infer the next action. Section V outlines our robotic system and execution framework. In Section VI, we show the results of the system, showcasing our system that integrates both pose and force outperforms the baseline systems. Finally, we conclude and discuss the results in Section VII.

II. Representation

We represent a task demonstrated by agents using an AOG consisting of: i) spatial knowledge to encode the poses of objects and manipulators, and ii) temporal knowledge to encode action sequencing.

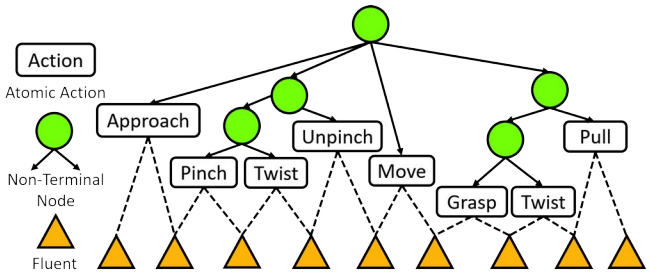


Fig. 2: An example of parse graph. Actions are executed in temporal order from left to right.

A. And-Or Graph (AOG)

An AOG is a graph-based grammar [8] encoding compositional variability in the demonstrated task sequences. Formally, an AOG $\mathcal{G}$ is represented by a 4-tuple:

$$\mathcal{G} = \langle U, V, \Delta, \Omega_F \rangle. \quad (1)$$

An And-node $u \in U$ represents a decomposition of the graph into sub-graphs, and an Or-node $v \in V$ acts as a switch among multiple alternate sub-configurations. The terminal nodes Δ is a set of sub-components representing the lowest level of resolution in the graph. Ω_F represents a set of attributes derived from the terminal nodes. In the context of opening bottles, $\Delta = \{a_1, \ldots, a_m\}$ corresponds to a set of *atomic actions* (Section II-B) executed during the task, and Ω_F is a set of *fluent functions* (Section II-C) that operate on terminal nodes.

A parse graph, denoted pg, is a specific parse of the AOG by selecting a sub-configuration at each Or-node in the graph. An example of a pg is shown in Fig. 2, simultaneously incorporating both spatial and temporal knowledge, where the spatial knowledge captures the physical configuration of the robot environment and fluents, and temporal knowledge encodes the sequence of atomic action to complete the task.

B. Atomic Actions

The concept of atomic actions [41] or action primitives [42] were proposed in the computer vision community. They are equivalent to the concept of movement primitives in robotics literature [43], [44] and represent the finest resolution of an action sequence. In this paper, both the human and robot actions are modeled using atomic actions. We aggregate each observed atomic action a_k^h from the demonstration to form the human dictionary of atomic action $\Delta_h = \{a_k^h\}$ and endow the robot with a dictionary of atomic actions, denoted $\Delta_r = \{a_k^r\}$. Here, the subscript k indicates the k-th atomic action in the action sequence. The correspondences between human and robot action labels were manually mapped. Each atomic action represents a 4D human-object interaction (4DHOI) unit, as in [45].

C. Fluents

From the human demonstrations, an auto-encoder is trained to embed the space of observed hand geometries, force distributions and the corresponding action label into a low-dimensional subspace. Changes in this low-dimensional subspace correspond to fluent changes. Each fluent function maps the high-dimensional scene configuration, s_k, to a real value, $f(s_k) \mapsto \mathbb{R}$. A fluent change represents a transition between two scene configurations, $\nabla f(s_i, s_j) = f(s_j) - f(s_i)$.

Fig. 3: Bottles used in experiments with different safety mechanism: (1) *push-and-twist*, (2) *pinch-and-twist*, (3) *push-and-twist*, and (4) *push-and-twist*. (5) Bottle with no safety mechanism.

For generality, we denote the action at step k as a_k, regardless of whether the action was performed by a human or robot. We denote the scene configuration of the pre-condition as s_k and the post-condition as s_{k+1}. Each action can be characterized by the changes it imposes across all fluents, denoted $\nabla f^{a_k} = \{\nabla f_i(s_k, s_{k+1}), \quad i = 1 \ldots n\}$.

Using this notion of fluent changes, the AOG encodes perceptual causality [46], represented by state changes between terminal nodes. We express this causal change as a structured equation model (SEM) [47]; *i.e.*, $f_{k+1} = g_{a_k}(f_k)$. This definition relies on the assumption that the human demonstrator/robot is the only causal agent in the environment and the *inertia action* assumption [48]. These two assumptions imply a perceptual causal chain between the agent's previous action and the next action; *i.e.*, the post-condition fluents of the previous action are the pre-condition fluents of the current action, depicted by the chain of fluents in Fig. 2.

III. Data Collection

A human demonstrator performed opening various types of bottles shown in Fig. 3. Some of the bottles contain child-safety locking mechanisms that require a procedure beyond simply twisting to unscrew the cap. Most child-safety locks require a particular force to be exerted on a particular part of the bottle. These forces are difficult to infer from visual observation alone. We collected human data on bottles 2, 3, and 5. The remaining bottles were reserved for testing.

a) Tactile Glove: We use a tactile glove [7] to capture these applied forces. The glove reconstructs the pose of each finger using IMUs and detects forces using Velostat sensors on the palm and phalanges. This glove provides 71 degrees of freedom including all pose and force measurements, resulting in an accurate model of the pose of the hand and the forces exerted by each phalange.

b) Experiment Setup: A Vicon motion capture system is used to record the ground truth of poses. The experimental setup is shown in Fig. 4. Fiducials are attached to each bottle and its lid to track the pose of object parts. One additional fiducial is attached to the back of the tactile glove to capture wrist pose in world space. A camera is used to record the video of data collection procedures to help label the ground truth later.

c) Data Collection: Approximately 10 trials are collected for each grasping strategy for each bottle. Examples are shown in Fig. 5. Bottle 2 only has one grasping strategy: *pinch-and-twist*. Bottle 3 has two different strategies: *push-and-twist* using the palm, or *push-and-twist* using fingers. Bottle 5 has three valid strategies because it lacks a safety mechanism: *twist*, *push-and-twist*, or *pinch-and-twist*.

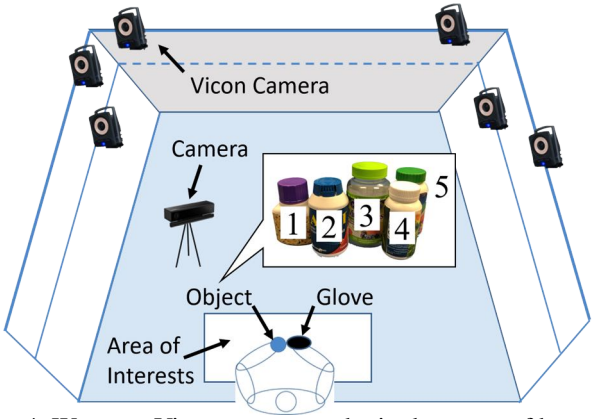


Fig. 4: We use a Vicon system to obtain the poses of human's wrist and object's parts. The camera is used to record the data collection procedure. The data is collected on bottles (2), (3) and (5), which require *pinch-and-twist*, *push-and-twist* and *twist* to open, respectively.

Each demonstration is manually labelled, mitigating the correspondence problem between a human action and a robot action. The timestamps of the labeling provide the transition boundaries between actions, *i.e.*, the post-condition of the labelled action and the pre-condition of the next action.

IV. Imitation Learning

A. Problem Definition

The planning objective is to find the best next action a^*_{k+1} given the observed partial parse graph $pg_k = (a_0, \dots, a_k)$. The pg is planned within the pre-defined action space, and fluents are used as observations. We plan this problem by minimizing the energy of the partial parse graph at each time step:

$$p(pg_{k+1}|pg_k, f_k) = \frac{1}{Z}\exp\{-\mathcal{E}(pg_{k+1}|pg_k, f_k)\}, \quad (2)$$

where $Z = \sum_{pg_{k+1}} \exp\{-\mathcal{E}(pg_{k+1}|pg_k, f_k)\}$ is the partition function. We decompose the energy of the parse graph into a top-down term and a bottom-up term, and adopt the notion of top-down and bottom-up as γ and β channels [49] of influence for inference in And-Or graphs, respectively. We define $\mathcal{E}(pg_{k+1}|pg_k, f_k)$ as

$$\mathcal{E}(pg_{k+1}|pg_k, f_k) = \mathcal{E}_\gamma(pg_{k+1}|pg_k) + \mathcal{E}_\beta(pg_{k+1}|pg_k, f_k), \quad (3)$$

$$\text{where} \quad \mathcal{E}_\gamma(pg_{k+1}|pg_k) = -\log\left[p(pg_{k+1}|pg_k))\right], \quad (4)$$

$$\mathcal{E}_\beta(pg_{k+1}|pg_k, f_k) = -\log\left[p(a_{k+1}|a_k, f_k)\right], \quad (5)$$

which incorporates two action planning mechanisms:

- **Top-down Term:** $p(pg_{k+1}|pg_k)$ plans the next action given the sequence of previous actions. It represents the *long-term* relation between all the previous actions and the next action. In this paper, an action grammar represented by AOG is first induced using all the valid action sequences. An Earley parser [50] is then adopted to parse the likelihood. See details in Section IV-B.
- **Bottom-up Term:** $p(a_{k+1}|a_k, f_k)$ plans the next action using both the current action label and observed fluent. This term encodes a *short-term* relation using the current fluent in addition to the pose and force pose sensing. In this paper, we convert this planning task to a classification problem, using a neural network to select the action with highest probability. See details in Section IV-C.

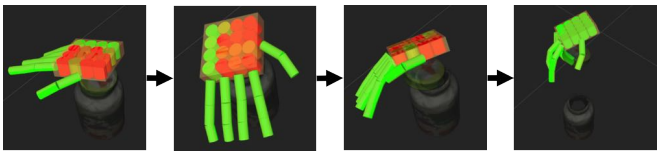


Fig. 5: The tactile glove computes the pose of human's phalanxes according to the pose of human's wrist and measure the force applied on human's hand.

B. Action Planning using AOG

a) AOG Induction: From labelled action sequences of human demonstration, an action grammar G represented by AOG is induced using method presented by Tu *et al.* [51], resulting in a stochastic context-free grammar with probabilistic Or-nodes. Examples are shown in Fig. 6. The objective function is the posterior probability of the grammar given the training data X:

$$p(G|X) \propto p(G)p(X|G) = \frac{1}{Z}e^{-\alpha||G||}\prod_{pg_i \in X} p(pg_i|G), \quad (6)$$

where $pg_i = (a_1, a_2, \dots, a_m) \in X$ represents a valid parse graph of atomic actions with length m from the demonstrator.

b) Top-down Parsing Likelihood: Given the learned AOG $\mathcal{G}$, for a grammatically complete parse graph $s = (a_0, \dots, a_K)$, the parsing likelihood is simply the Viterbi likelihood, denoted by $p(s)$. For an incomplete parse $pg_k = (a_0, \dots, a_k)$ with length of $k < K$, the parsing likelihood is given by the sum over all grammatically possible actions sequences that begin with pg_k:

$$p(pg_k) = \sum_{s \in \mathcal{G}, s_k = pg_k} p(s), \quad (7)$$

where pg_k denotes the first k actions in the parse graph pg. By computing $p(pg_{k+1})$ and $p(pg_k)$ using the Earley parsing likelihood, we compute the top-down term, $p(pg_{k+1}|pg_k)$, through Bayes' rule. The top-down term encodes long-range temporal constraints induced by the AOG.

C. Action Planning using Fluents

We use tactile glove measurements and haptic feedback signals to learn: i) a low-dimensional embedding of the human demonstration, ii) a bottom-up term to plan the next action based on the low-dimensional human embedding, and iii) an embodiment mapping between the robot and the low-dimensional human embedding.

a) Low-dimensional Embedding: We use an auto-encoder to encode the scene configuration into a low-dimensional representation as fluents (Fig. 7(a)). Changes inside this subspace are treated as fluent changes and are used to infer the next action with observed haptic feedback from the robot. Within this subspace, we train a bottom-up term, $p(a_{k+1}|a_k, f_k)$, to plan the next action using haptic observations of the post-condition of the previous action.

The contact force and pose measurements from the tactile glove are reoriented to the reference frame of the wrist, and concatenated into a feature vector with 159 dimensions. An encoder-decoder architecture, illustrated in Fig. 7(a), is used to learn a 8-dimensional embedding and reconstructs the full feature from this embedding under a criterion that minimizes the squared residuals between the original feature and the

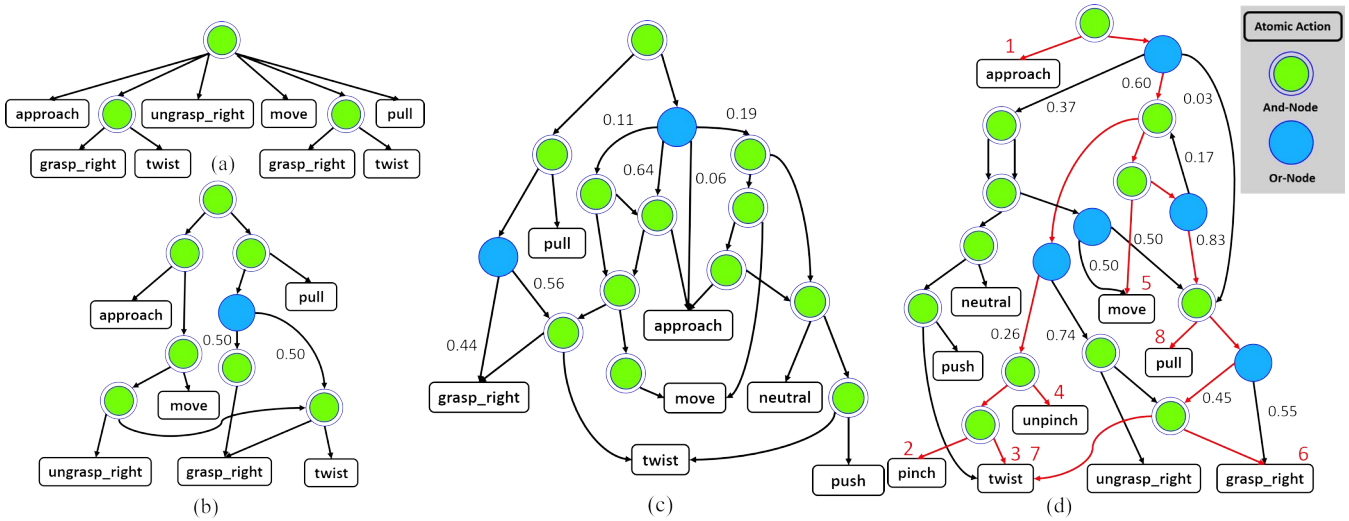


Fig. 6: AOG induced from human demonstrations using 1 example (a), 5 examples (b), 36 examples (c), and 65 examples (d). (d) also shows Fig. 2 parsed in an AOG, highlighted in red. Numbers indicate temporal ordering of atomic actions.

reconstruction:

$$l(\theta;\mathbf{x}^h) = \frac{1}{N}\sum_{i=1}^{N}(x_i^h - \psi(x_i^h;\theta))^2, \tag{8}$$

where x_i^h represents one of the N human demonstrations and $\psi(x_i;\theta)$ represents the reconstruction.

b) Bottom-up Action Planning: The bottom-up term $p(a_{k+1}|a_k, f_k)$ takes the form of a multi-class classifier to plan one of the 13 output actions (Fig. 7(b)). This classification network takes its input from the embedding layer of the auto-encoder and a one-hot encoding of the current action. A softmax layer is used to interpret it as a probability distribution, and the network is trained by minimizing the normalized cross-entropy. All internal layers are linear matrix operators, and use sigmoids for their non-linearities. Combined with the low-dimensional embedding, the bottom-up term incorporates raw tactile signals during manipulations, thus complementing the top down constraints from the action grammar parsing.

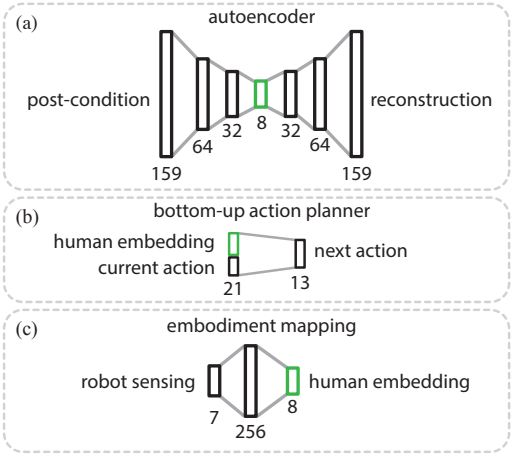


Fig. 7: (a) Autoencoder to project human demonstration into low-dimensional subspace. (b) Classifier used to plan the next action using a low-dimensional embedding of human tactile feedback. (c) Embodiment mapping used to map robot states to equivalent human demonstration states. Each rectangle represents a vector, and each corresponding number is the length of the vector. The green rectangle represents the low-dimensional human embedding vector.

c) Embodiment Mapping: The embodiment mapping seeks a function $s_h = \hat{\phi}(s_r)$, where s_h represent the human state of the demonstration and s_r represents the robot's state during execution (Fig. 7(c)). This function maps haptic sensing on the robot to the low-dimensional embedding of tactile measurements from the human demonstration. A neural network is trained to approximate this function using a small number of robot examples (approximately 15 examples). We supervise robot executions sampled from the learned AOG using the robot's dictionary Δ_r to ensure only successful robot states are mapped to successful demonstrator states. The loss function for this network is the squared residuals:

$$l(\theta;\mathbf{x}^h,\mathbf{x}^r) = \frac{1}{N}\sum_{i=1}^{N}(\phi(x_i^h) - \hat{\phi}(x_i^r;\theta))^2, \tag{9}$$

where $\mathbf{x}^h$ represents human states, $\mathbf{x}^r$ represents equivalent robot states, ϕ represents the low-dimensional embedding of human data, and $\hat{\phi}$ represents the embodiment mapping function. The robot utilizes this mapping to plan the next action using the bottom-up term: first map its state to an equivalent human state, then use the human state to plan which action to execute using the bottom-up action planner.

V. Implementation

A. Robot Platform Setup

We use a dual-armed 7-DoF Baxter robot from Rethink Robotics mounted on a DataSpeed Mobility Base as our robot platform. The robot is equipped with a ReFlex TackkTile gripper on the right wrist, and a Robotiq S85 parallel gripper on the left. In addition, we use Simtrack [52] for object pose estimation and tracking with a Kinect One sensor. The entire system runs on ROS [53], and arm motion planning is computed using MoveIt! [54]. For object grasping, we implement a geometry based grasping planner to generate grasping poses from CAD models of the objects.

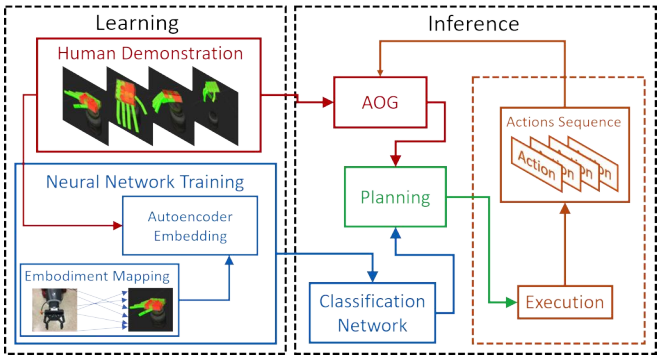


Fig. 8: System architecture. Blue: action planning using fluents as a bottom-up process. Red: action planning using AOG as a top-down process. Green: action planning. Brown: robot execution.

B. System Architecture

The system architecture consists of three major components shown in Fig. 8:

- **Learning:** The learning phase includes a top-down process and a bottom-up process. The top-down representation is built from segmented human demonstrations, and an AOG is induced to represent valid action sequences (see Section IV-B). To learn the bottom-up knowledge, three neural networks are trained from raw sensor data (see Section IV-C).
- **Inference:** During the inference, the top-down term is computed by the Earley parser. The embodiment mapping and classification network are used to compute the bottom-up term, as outlined in Section IV-A. We plan the next action using Equation 2 with the corresponding top-down and bottom-up terms.
- **Execution:** Robot executes the next action either by sampling the AOG, using haptic feedback, or both according to Equation 2.

VI. Experiments and Results

A. Experiment Setup

Five bottles were used in the evaluation as shown in Fig. 3. Bottles 2, 3, and 5 were used during data collection, while the remaining bottles were reserved for testing. Bottles 1, 2, 3, and 4 all have safety mechanisms while bottle 5 does not.

An action sequence is deemed successful if the robot opens the bottle; otherwise, the sequence is a failure. If the robot opens the bottle before finishing the sampled execution, we consider the action sequence that it performed is correct and discard remain actions. We conducted over 300 opening experiments over all of the bottles, resulting in three groups of quantitative results. Each bottle was tested approximately 60 times.

B. Evaluation Criteria

While there may be multiple ways to open each bottle, not all methods are considered equivalent. For instance, Bottle 5 has no safety mechanism, so while *push-and-twist* and *pinch-and-twist* may succeed in opening bottle 5, there is no reason to execute anything other than *twist*. This distinction naturally leads to two levels of evaluation criteria: i) by the end results only, *i.e.*, whether a sequence of actions can successfully open a bottle, and ii) not only successfully open a bottle but also efficiently.

As illustrated above, human demonstrator is treated as an oracle and the corresponding action sequences as perfect executions. We separate robot executions into four different categories:

1) Success, where the robot successfully executed an action sequence that is an exact match to one of the sequences from the human demonstrator;
2) Success, but using at least one extra or wrong action;
3) Failure due to using the wrong action sequence; and
4) Failure due to improper execution (*e.g.* low motor execution accuracy or grasping failure).

C. Qualitative and Quantitative Results

For qualitative analysis, Fig. 9 shows the robot successfully opening two bottles with (Fig. 9(a)) and without (Fig. 9(b)) pushing the bottle lid. The force-torque sensor readings reflect distinguishable differences between performing *push-and-twist* (Fig. 9(c)) and *twist* (Fig. 9(d)).

We set up three groups of experiments for quantitative results analysis. Table I shows the results of using top-down only planning, in which the robot executes a sampled action sequence only from the AOG. This method describes the order in which actions were executed but does not capture haptics during manipulations.

Table II shows the results of using bottom-up only planning. This method incorporates the haptic feedback from the robot sensing, but lacks long-term temporal constraints from the AOG, *i.e.*, it executes a Markovian planning process, in which the next action is determined by the previous action and the current observations as outlined in Section IV.

Table III shows the results of integrating both the top-down planning provided by the AOG and the bottom-up haptic feedback. By utilizing both terms, the temporal sequence of actions is not generated only by sampling from the AOG; instead, each action is generated sequentially by minimizing Equation 2.

The proposed top-down and bottom-up planning (Table III) yields large performance improvements over either

TABLE I: Baseline 1, top-down only planning

Evaluation	bot. 1	bot. 2	bot. 3	bot. 4	bot. 5
Success	8.7%	5.6%	4.4%	8.7%	26.1%
Success (extra/wrong)	21.7%	5.6%	34.8%	47.8%	39.1%
Failure (action)	69.6%	77.7%	60.8%	34.8%	30.4%
Failure (execution)	0%	11.1%	0%	8.7%	4.4%

TABLE II: Baseline 2, bottom-up only planning

Evaluation	bot. 1	bot. 2	bot. 3	bot. 4	bot. 5
Success	4.4%	0%	4.4%	0%	4.4%
Success (extra/wrong)	13%	11.8%	30.4%	42.9%	17.4%
Failure (action)	82.6%	76.4%	65.2%	57.1%	78.2%
Failure (execution)	0%	11.8%	0%	0%	0%

TABLE III: Proposed, top-down and bottom-up planning

Evaluation	bot. 1	bot. 2	bot. 3	bot. 4	bot. 5
Success	8.7%	17.6%	17.4%	20%	60.9%
Success (extra/wrong)	52.2%	17.6%	65.2%	73.3%	17.4%
Failure (action)	39.1%	64.8%	13%	6.7%	21.7%
Failure (execution)	0%	0%	4.4%	0%	0%

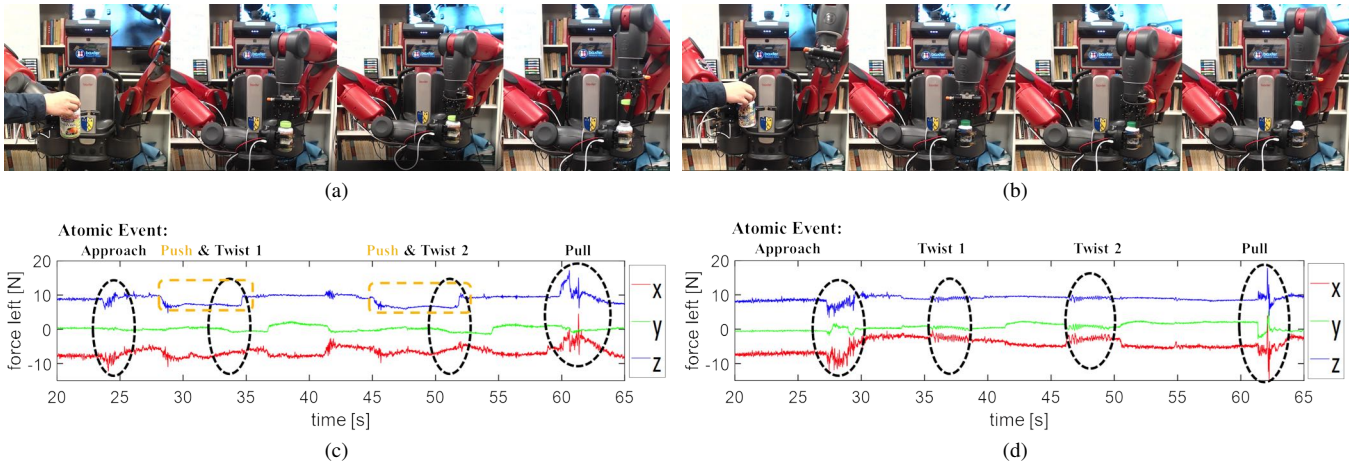


Fig. 9: (a) Robot opening bottle 3, showing actions *approach*, *push*, *twist*, and *pull* from left to right. (b) Robot opening bottle 5, showing actions *approach*, *grasp*, *twist*, and *pull*. Force-torque sensor readings while opening bottle 3 (c) and bottle 5 (d), showing clear, distinguishable differences from raw sensor data.

the top-down (Table I) or bottom-up (Table II) only method. The rate of Success and Success with extra/wrong are dramatically improved while the failure rate due to wrong actions sequences drops significantly.

D. Discussion

a) Why it is important to integrate both top-down and bottom-up terms? In our proposed method, top-down planning generates an action from the non-Markovian AOG, while the bottom-up planning formulates a Markovian process according to robot's haptic feedback. These two processes are complementary to each other and crucial to correctly executing a manipulation task. Specifically, i) the top-down term represents the structure of the task, generating the next action based on previous semantic knowledge and preventing executing irrelevant actions. ii) The bottom-up term encodes real-time sensing information, capturing subtle interactions during manipulations. By combining these two terms, our method is capable of learning from small examples of human demonstrations and planning actions on the fly based on task structure and real-time haptic sensing.

b) Why the success rate of bottle 2 is low? The robot has no haptic feedback and geometry information prior to touching the bottle with its gripper. By sampling the first action after *approach* from the AOG, the probability to plan *pinch* is around 15%, due to the frequency in the human demonstrations. While not reported in Table III, the perfect successful rate for bottle 2 is 100% if the first action after *approach* is *pinch*. Other work [55] has augmented AOG nodes with attributes to turn the AOG into a context-sensitive grammar. A context-sensitive grammar would increase perfect success rates by considering the type of bottle directly in the top-down term, rather than our current method implicitly inferring the bottle type from haptic feedback.

c) Can the robot derive novel manipulations that are not presented in human demonstrations? In our opinion, there are at least two types of novel manipulations that a robot can derive from human demonstrations: i) generating new action sequences, and ii) generating new actions. In this paper, the proposed method demonstrates the capability of generating novel action sequences through a compositional grammar. However, generating new actions is much more difficult, as the structure and capability of human hands and robot grippers could be dramatically different. For instance, a human demonstration may need to twist twice to open a bottle lid, while a robot gripper may only need to twist once, since some robot grippers are capable of rotating with more freedom than human wrist. Such differences lead to the different success rates of bottle 1, 3, and 4 even though they all require *push-and-twist*: bottle 1 must *push-and-twist* at least twice to open, while bottles 3 and 4 require only one *push-and-twist* action. If the robot could learn and infer the degree of rotation required to open the bottle, the robot could generate a new action to achieve tasks. However, the proposed method does not explore the parameterization of each atomic action in the presented work.

VII. Conclusion

In this work, we present a novel method of naturally capturing visually hidden states of a task and transferring them to the robot through human demonstrations using a tactile glove. The tactile glove provides a data collection method to capture visually hidden causal changes in the scene. Using this latent encoding of the scene, we learn a model to plan the actions of the human demonstrator. The human demonstrations are used to induce an AOG, and the AOG is used to supervise successful executions of opening a bottle.

The robot states of successful executions are mapped to successful demonstrations from the human demonstrator using a low-dimensional embedding of the human tactile feedback. This embodiment mapping solves the correspondence problem using a relatively small number of supervised robot executions. The robot utilizes this mapping in conjunction with the top-down and bottom-up terms to infer the next action to execute.

The proposed method (Table III) shows a marked improvement over two baselines (Table I and II), demonstrating the top-down and bottom-up terms work together to increase the success rate in comparison to using either method alone.

Future Work

This work paves the way for additional work regarding visually latent states and corresponding embodiment mappings. We would like to investigate methods to make the system less supervised by clustering the human demonstrations. From the clusters, the robot may not possess an equivalent action in its dictionary and may need to search its action space for an action with equivalent pre- and post-conditions.

The framework presented here could be used to attempt functionally equivalent tasks [4]. In this way, the robot could demonstrate understanding the dynamics of the task that needs to be replicated and which can be safely ignored. Experimenting to find functional equivalence is closely related to counterfactual reasoning in the causal domain; such explorations establish causal connections between actions and their effects.

Acknowledgement We thank Ruiqi Gao of UCLA Statistics Department and Shu Wang of Fudan University Electrical Engineering Department for assistance on experiments. The work reported herein was supported by DARPA XAI grant N66001-17-2-4029, DARPA SIMPLEX grant N66001-15-C-4035 and ONR MURI grant N00014-16-1-2007.

REFERENCES

[1] A. J. Schmid, N. Gorges, D. Goger, and H. Worn, "Opening a door with a humanoid robot using multi-sensory tactile feedback," in *ICRA*, IEEE, 2008.
[2] W. Zhao, J. Zhang, J. Min, and J. Chai, "Robust realtime physics-based motion control for human grasping," *TOG*, vol. 32, no. 6, p. 207, 2013.
[3] Y. Wang, J. Min, J. Zhang, Y. Liu, F. Xu, Q. Dai, and J. Chai, "Video-based hand manipulation capture through composite motion control," *TOG*, vol. 32, no. 4, p. 43, 2013.
[4] Y. Zhu, Y. Zhao, and S.-C. Zhu, "Understanding tools: Task-oriented object modeling, learning and recognition," in *CVPR*, 2015.
[5] T.-H. Pham, A. Kheddar, A. Qammaz, and A. A. Argyros, "Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces," in *CVPR*, 2015.
[6] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S.-C. Zhu, "Inferring forces and learning human utilities from videos," in *CVPR*, 2016.
[7] H. Liu, X. Xie, M. Millar, M. Edmonds, F. Gao, Y. Zhu, V. J. Santos, B. Rothrock, and S.-C. Zhu, "A glove-based system for studying hand-object manipulation via joint pose and force sensing," in *IROS*, IEEE, 2017.
[8] S.-C. Zhu, D. Mumford, *et al.*, "A stochastic grammar of images," *Foundations and Trends in Computer Graphics and Vision*, vol. 2, no. 4, pp. 259–362, 2007.
[9] I. Newton and J. Colson, *The Method of Fluxions and Infinite Series; with Its Application to the Geometry of Curve-lines*. 1736.
[10] L. Dipietro, A. M. Sabatini, and P. Dario, "A survey of glove-based systems and their applications," *Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 4, pp. 461–482, 2008.
[11] T. Taylor, S. Ko, C. Mastrangelo, and S. J. M. Bamberg, "Forward kinematics using imu on-body sensor network for mobile analysis of human kinematics," in *Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2013.
[12] H. G. Kortier, V. I. Sluiter, D. Roetenberg, and P. H. Veltink, "Assessment of hand kinematics using inertial and magnetic sensors," *Journal of Neuroengineering and Rehabilitation*, vol. 11, no. 1, p. 70, 2014.
[13] G. Santaera, E. Luberto, A. Serio, M. Gabiccini, and A. Bicchi, "Low-cost, fast and accurate reconstruction of robotic and human postures via imu measurements," in *ICRA*, IEEE, 2015.
[14] N. S. Kamel, S. Sayeed, and G. A. Ellis, "Glove-based approach to online signature verification," *TPAMI*, vol. 30, no. 6, pp. 1109–1113, 2008.
[15] R. K. Kramer, C. Majidi, R. Sahai, and R. J. Wood, "Soft curvature sensors for joint angle proprioception," in *IROS*, IEEE, 2011.
[16] Y. Gu, W. Sheng, M. Liu, and Y. Ou, "Fine manipulative action recognition through sensor fusion," in *IROS*, IEEE, 2015.
[17] E. Jeong, J. Lee, and D. Kim, "Finger-gesture recognition glove using velostat," in *ICCAS*, IEEE, 2011.
[18] J. Low, P. Khin, and C. Yeow, "A pressure-redistributing insole using soft sensors and actuators," in *IROS*, IEEE, 2015.
[19] G. Pugach, A. Melnyk, O. Tolochko, A. Pitti, and P. Gaussier, "Touch-based admittance control of a robotic arm using neural learning of an artificial skin," in *IROS*, IEEE, 2016.
[20] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009.
[21] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.
[22] J. Lieberman and C. Breazeal, "Improvements on action parsing and action interpolation for learning through demonstration," in *International Conference on Humanoid Robots*, IEEE, 2004.
[23] P. Kormushev, D. N. Nenchev, S. Calinon, and D. G. Caldwell, "Upper-body kinesthetic teaching of a free-standing humanoid robot," in *ICRA*, IEEE, 2011.
[24] C. L. Campbell, R. A. Peters, R. E. Bodenheimer, W. J. Bluethmann, E. Huber, and R. O. Ambrose, "Superpositioning of behaviors learned through teleoperation," *T-RO*, vol. 22, no. 1, pp. 79–91, 2006.
[25] K. Kukliński, K. Fischer, I. Marhenke, F. Kirstein, V. Maria, N. Krüger, T. R. Savarimuthu, *et al.*, "Teleoperation for learning by demonstration: Data glove versus object manipulation for intuitive robot control," in *ICUMT*, IEEE, 2014.
[26] P. Kormushev, S. Calinon, and D. G. Caldwell, "Imitation learning of positional and force skills demonstrated via kinesthetic teaching and haptic input," *Advanced Robotics*, vol. 25, no. 5, pp. 581–603, 2011.
[27] G. M. Hayes and J. Demiris, *A robot controller using learning by imitation*. University of Edinburgh, Department of Artificial Intelligence, 1994.
[28] M. Muhlig, M. Gienger, S. Hellbach, J. J. Steil, and C. Goerick, "Task-level imitation learning using variance-based movement optimization," in *ICRA*, IEEE, 2009.
[29] S. Ross, G. J. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *AISTATS*, pp. 627–635, 2011.
[30] M. Falahi and M. Jannatifar, "Using orthogonal basis functions and template matching to learn whiteboard cleaning task by imitation," in *ICCKE*, IEEE, 2013.
[31] J. D. Langsfeld, K. N. Kaipa, R. J. Gentili, J. A. Reggia, and S. K. Gupta, "Incorporating failure-to-success transitions in imitation learning for a dynamic pouring task," in *Workshop on Compliant Manipulation: Challenges and Control, Chicago, IL*, 2014.
[32] C. Paxton, F. Jonathan, M. Kobilarov, and G. D. Hager, "Do what i want, not what i did: Imitation of skills by planning sequences of actions," in *IROS*, IEEE, 2016.
[33] C. Xiong, N. Shukla, W. Xiong, and S.-C. Zhu, "Robot learning with a spatial, temporal, and causal and-or graph," in *ICRA*, IEEE, 2016.
[34] T. Shu, X. Gao, M. Ryoo, and S.-C. Zhu, "Learning social affordance grammar from videos: Transferring human interactions to human-robot interactions," in *ICRA*, 2017.
[35] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *ICML*, p. 1, ACM, 2004.
[36] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning," in *IJCAI*, vol. 51, pp. 1–4, 2007.
[37] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning.," in *AAAI*, vol. 8, pp. 1433–1438, Chicago, IL, USA, 2008.
[38] B. Huang, M. Li, R. L. De Souza, J. J. Bryson, and A. Billard, "A modular approach to learning manipulation strategies from human demonstration," *Autonomous Robots*, vol. 40, no. 5, pp. 903–927, 2016.
[39] J. Sung, J. K. Salisbury, and A. Saxena, "Learning to represent haptic feedback for partially-observable tasks," *ICRA*, 2017.
[40] Y. Derimis and G. Hayes, "Imitations as a dual-route process featuring predictive and learning components: a biologically plausible computational model," *Imitation in animals and artifacts*, pp. 327–361, 2002.
[41] M. Pei, Z. Si, B. Z. Yao, and S.-C. Zhu, "Learning and parsing video events with goal and intent prediction," *CVIU*, vol. 117, no. 10, pp. 1369–1383, 2013.
[42] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," in *CVPR*, IEEE, 2008.
[43] S. Schaal, J. Peters, J. Nakanishi, and A. Ijspeert, "Learning movement primitives," in *Robotics Research. The Eleventh International Symposium*, pp. 561–572, Springer, 2005.
[44] A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann, "Probabilistic movement primitives," in *NIPS*, 2013.
[45] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for event and object recognition," in *ICCV*, 2013.
[46] B. J. Scholl and P. D. Tremoulet, "Perceptual causality and animacy," *Trends in cognitive sciences*, vol. 4, no. 8, pp. 299–309, 2000.
[47] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 2009.
[48] N. McCain, H. Turner, *et al.*, "Causal theories of action and change," in *AAAI*, 1997.
[49] T. Wu and S.-C. Zhu, "A numerical study of the bottom-up and top-down inference processes in and-or graphs," *IJCV*, vol. 93, no. 2, pp. 226–252, 2011.
[50] J. Earley, "An efficient context-free parsing algorithm," *Communications of the ACM*, vol. 13, no. 2, pp. 94–102, 1970.
[51] K. Tu, M. Pavlovskaia, and S.-C. Zhu, "Unsupervised structure learning of stochastic and-or grammars," in *NIPS*, 2013.
[52] K. Pauwels and D. Kragic, "Simtrack: A simulation-based framework for scalable real-time object pose detection and tracking," in *IROS*, IEEE, 2015.
[53] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, 2009.
[54] I. A. Sucan and S. Chitta, "Moveit!," *Online at http://moveit. ros. org*, 2013.
[55] S. Park, B. X. Nie, and S.-C. Zhu, "Attribute and-or grammar for joint parsing of human attributes, part and pose," *ICCV*, 2015.