2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Unsupervised Learning of Hierarchical Models for Hand-Object Interactions

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

***Abstract*— Contact forces of the hand are visually unobservable, but play a crucial role in understanding hand-object interactions. In this paper, we propose an unsupervised learning approach for manipulation event segmentation and manipulation event parsing. The proposed framework incorporates hand pose kinematics and contact forces using a low-cost easy-to-replicate tactile glove. We use a temporal grammar model to capture the hierarchical structure of events, integrating extracted force vectors from the raw sensory input of poses and forces. The temporal grammar is represented as a temporal And-Or graph (T-AOG), which can be induced in an unsupervised manner. We obtain the event labeling sequences by measuring the similarity between segments using the Dynamic Time Alignment Kernel (DTAK). Experimental results show that our method achieves high accuracy in manipulation event segmentation, recognition and parsing by utilizing both pose and force data.**

I. Introduction

Consider a complex manipulation event of a person opening a medicine bottle with safety lock (Fig. 1). During this process, a number of movement primitives were performed: *grasp*, *push-and-twist*, *push-and-twist*, *twist*, and finally *pull* the lid off the bottle. Even with the most state-of-the-art action understanding and recognition algorithms (see survey [1], [2]), it is still challenging to segment such action sequence and parse the manipulation event. This is due to three major difficulties: i) severe occlusions happen during fine manipulation, especially self-occlusions, ii) in subtle manipulation tasks, visual data may not be able to reveal adequate knowledge to capture the quintessence. Certain actions are hard to detect using skeleton data alone but need additional force readings *e.g.*, whether an action of pushing was performed during twisting the lid, and iii) ground truth data is difficult to obtain using vision sensor alone, oftentimes impossible to obtain the needed information (*e.g.*, the force readings, and accurate finger poses during occlusions).

In this paper, we present an unsupervised learning method for manipulation event segmentation, recognition and parsing. The method not only accounts for the aforementioned challenges, but also captures the temporal hierarchical structure of the manipulation sequence using a grammar model—a temporal And-Or graph (T-AOG). Specifically, we investigate the manipulation actions of opening different types of medicine bottles. Some examples are shown in Fig. 4a.

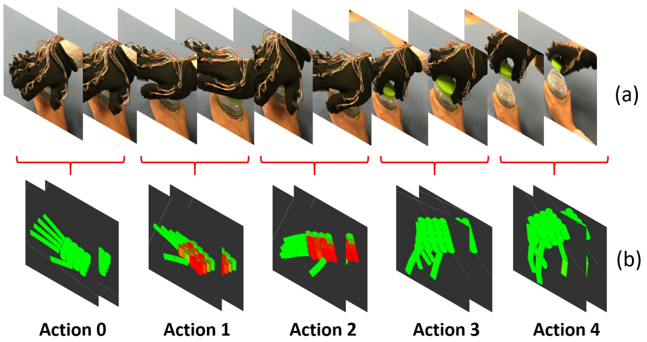


Fig. 1: (a) A sequence of movement primitive demonstrated by an agent for a manipulation task–opening a medicine bottle captured by a tactile glove. (b) Reconstructed force and pose data using the tactile glove. Our purposed method segments and parses the noisy inputs of force and pose in an unsupervised fashion.

Bottle 1 has no safety lock and can be opened by simply twisting the lid. *Bottle 2* requires pressing the lid while twisting. Pinching the safety lock is needed to open *Bottle 3*. Importantly, some actions (*e.g.*, pressing, pinching) are difficult to observe visually, thus require additional sensing for action recognition.

To obtain the force readings during manipulations, we propose to study hand-object interactions with additional force information through a low-cost, easy-to-replicate tactile glove [3]. Although some efforts have been shown to recover the forces during interactions using vision-based methods [4], [5], [6], [7], [8], it remains an open problem without adopting a hardware-based solution. Using a tactile glove can reliably retrieve contact forces to overcome the limitation of using visual data alone.

By observing the data collected using the tactile glove, such as the force exerted on the palm, we can learn that a push-down action is performed as well as a set of motion primitives that can best describe the action sequences. Thus, our system is able to “see”, in numerical terms, the forces during hand-object interactions. We argue that this is an important step in recognizing manipulation actions with visually latent force information.

Still, it is nearly impossible to understand and transfer the raw data (poses and forces) retrieved from the tactile glove *directly* to a robot due to different embodiments. Therefore, we need to reconstruct the semantic meanings of manipulation events from the human demonstration, allowing the transfer of abstract knowledge to a robot.

To recover the semantic meaning and model the temporal

[⋆] Xu Xie and Hangxin Liu contributed equally to this work.

[1] Xu Xie, Hangxin Liu, Mark Edmonds, Feng Gao, Siyuan Qi, Yixin Zhu and Song-Chun Zhu are with UCLA Center for Vision, Cognition, Learning, and Autonomy at Statistics Department. {xiexu, hx.liu, markedmonds, f.gao, syqi, yixin.zhu}@ucla.edu, sczhu@stat.ucla.edu.

[2] Brandon Rothrock is with the Jet Propulsion Laboratory, California Institute of Technology. rothrock@jpl.nasa.gov.

978-1-5386-3080-8/18/$31.00 ©2018 IEEE

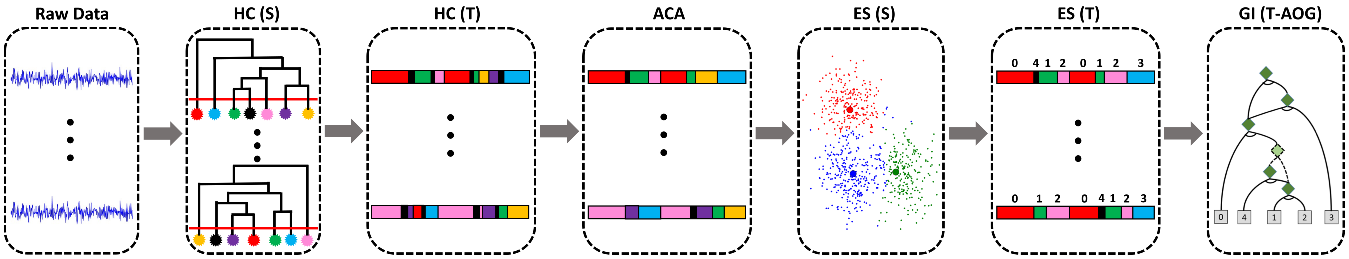


Fig. 2: Unsupervised learning pipeline of hand-object motion recognition. After collecting the raw data using a tactile glove, a spatial (HC (S)) and temporal (HC (T)) hierarchical clustering is performed on both force and pose data. An aligned cluster analysis (ACA) is adopted to further reduce the noise. Event segmentation (ES (S) and ES (T)) is achieved by merging motion primitives based on the distance measured by DTAK. Finally, a grammar is induced (GI) based on the segmented events, forming a T-AOG.

structure of actions in a hand-object interaction, we represent the manipulation sequence using a T-AOG, a temporal grammar model that captures the hierarchical structure of the action sequences. Its terminal nodes are motion primitives, *e.g.*, twisting and pressing, which is learned by unsupervised clustering over extracted features of the pose and force sensory inputs. To evaluate the effectiveness of our model, we compare the segmentation and labeling results of different sensory data with several baseline methods.

A. *Related Work*

a) Action Recognition: A number of approaches have been proposed for action recognition in various applications. This literature is too wide to survey here; we refer readers to two recent surveys for recognizing and parsing human actions [1], [2]. Recently, due to additional sensory input, RGB-D sensors such as Kinect are capable of estimating 3D poses from a single image [9]. Further studies have demonstrated impressive results of pose estimation and action recognition from RGB-D videos [10], [11], [12], [13], [14], [15]. These works, however, focuses on body-size action recognition without force sensing. In contrast, the presented work addresses the hand-size finer-grained manipulation actions with reconstructed forces.

b) Vision-based Force Estimation: Brubaker *et al.* estimated contact forces and internal joint torques using a mass-spring system [16], [17], [18]. More recently, Zhu *et al.* [6] and Pham *et al.* [7] proposed to use numerical differentiation methods to estimate hand-object interactions during manipulation tasks. In computer graphics, sophisticated physics-based soft-body simulation can calculate contact force from video [5], [4]. These work, however, requires prior knowledge of geometry and physical properties of the manipulated objects. By using a tactile glove, estimating forces in the present study does not rely on such assumptions.

c) Learning from Demonstration (LfD): A robot must recognize and understand the actions sufficiently in order to imitate the tasks from the demonstrations. LfD (also imitation learning, learning by watching, or apprenticeship learning) is too expansive to survey here; we refer readers to a survey [19]. In the last few years, with the recent rise of Convolutions Neural Networks, there are increasing interests in providing and parsing demonstrations using pure visual data [20] by learning action plans [21] and physical interactions [22] in complex and higher-level tasks, *e.g.*, cloth folding [23]. However, it is yet still difficult to convey force information from vision-based methods reliably.

d) Kinesthetic Teaching and Teleoperation: To address the above issue, the robotics community has been developing kinesthetic teaching or teleoperation approaches to recognize low-level motion primitives during hand-object interactions. These approaches are capable of transferring certain rich physical information such as force knowledge to robots. Manschitz *et al.* [24] presented a method to teach robots to unscrew a light bulb by moving primitives, which are represented by sequences of graphs. A more recent work was presented in [25]. Chebotar *et al.* [26] used spectral clustering and PCA to reduce the dimensionality in learning tactile feedback during performing scraping task. More challenging hand-object interaction tasks involving the manipulation of deformable objects were discussed using a similar approach [27]. Learning impedance behaviors and trajectory following skills was presented in [28] by combining robot's dynamical system and stiffness estimation.

B. *Contributions*

This paper makes three contributions:

1) We incorporate *invisible* force in addition to the conventional pose-based methods for event segmentation and parsing during fine-grained manipulation tasks. We show in the experiment that a better performance of motion recognition is achieved by jointly considering hand pose and force data.
2) We propose an unsupervised learning framework to learn a temporal grammar model (T-AOG) for hand-object interactions. The framework incorporates automatic clustering, segmentation, labeling, and high-level grammar induction. The grammar structure is shown to significantly improve the action recognition results compared to using clustering method alone.
3) We introduce a general method for modeling noisy and heterogeneous sensory data of hand-object manipulation.

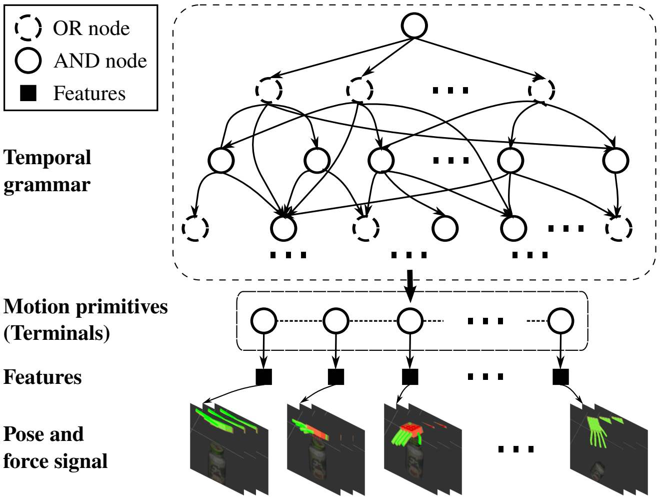


Fig. 3: Illustration of the T-AOG. The T-AOG is a temporal grammar in which the terminal nodes are motion primitives of hand-object interactions.

C. Overview

The remainder of this paper is organized as follows. In Section II, we introduce the representation T-AOG. In Section III, we present the learning algorithm consisting of hierarchical clustering and grammar induction. The inference algorithm of motion recognition is introduced in Section IV. In Section V, we demonstrate the data with additional force sensing indeed outperforms the data with either pose or force data only. Furthermore, our analysis shows the parsing results using T-AOG help improve the performance significantly compared with using clustering only.

II. Representation

We introduce a structural grammar model *Temporal And-Or Graph (T-AOG)* [29] to represent the temporal structure of a task. An AOG is a directed graph which describes a stochastic context free grammar (SCFG), providing a hierarchical and compositional representation. Formally, the AOG is defined as a five-tuple $G=(S,V,R,P,\Sigma)$, where S is a start symbol; V is a set of nodes which includes the non-terminal nodes V^{NT} and terminal nodes V^T: $V=V^{NT}\cup V^T$; $R=\{r:\alpha\to\beta\}$ is a set of production rules that represent the top-down sampling process from a parent node α to its child nodes β; $P:p(r)=p(\beta|\alpha)$ is the probability for each production rule; Σ is the language defined by the grammar, *i.e.*, the set of all valid sentences given the grammar.

In an AOG, the **non-terminal** nodes can be divided into two types: $V^{NT}=V^{AND}\cup V^{OR}$. An **And-node** is used to represent 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** is used to represent alternative configurations. 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 represent the entities that are not further decomposed or have different configurations. A **parse graph** pg is an instance of the AOG, where the And-nodes are decomposed and one of the child nodes is selected for the Or-nodes.

In particular, a T-AOG represents a set of all possible sequences to execute a certain task. The start node S represents an event category (*e.g.*, opening a bottle). The terminal nodes V^T represents the set of motion primitives that a human or a robot can perform (*e.g.*, approaching, twisting). An And-node is decomposed into sub-events or motion primitives as its child nodes. An Or-node encodes alternative solutions to perform a sub-task. A pg for an event is a sub-graph of T-AOG that captures the temporal structure of the scenario.

As shown in Fig. 3, features are extracted from the raw input sensory data and further segmented for semantic parsing. Pose and force features Γ are extracted based on a raw sensory input sequence I in time interval $[1,T]$. Each frame is labeled with motion primitive a_t. Aggregating together, we obtain a label sequence $A=\{a_t\}$. The segmentation of the sensory input sequence is defined as $\mathcal{T}=\{\gamma_k\}, k=1,\cdots,K$, where $\gamma_k=[t_k^1,t_k^2]$ represents a time interval in which the motion primitive remains the same. Later in this paper, we use a_{γ_k} to denote the motion label for the segment I_{γ_k}.

III. Learning of Hand-Object Interactions

The unsupervised learning pipeline is illustrated in Fig. 2. Given training sequences of raw sensory input of poses and forces, our goal of learning is to unsupervisedly learn i) the motion primitives in the sequences of hand-object interactions, ii) the event segmentation in every sequence, and iii) the high-level grammar structure (T-AOG) that captures every observed sequences of the hand-object interactions.

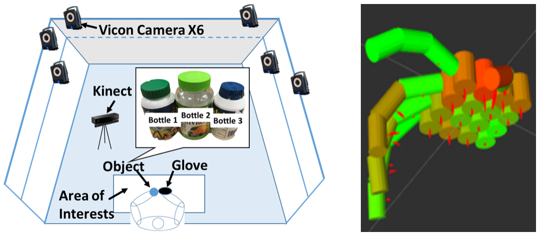
A. Unsupervised Learning of Motion Primitives

To recognize motion primitives of hand-object interactions, we adopt the agglomerative hierarchical clustering, capable of successively merging the similar features from the low-level features, without knowing the exact number of clusters in advance. The Wards agglomerative method is used to determine whether a merge is needed in each iteration:

$$\begin{aligned}\triangle(A,B) = & \sum_{i\in A\cup B}||\vec{x}_i-\vec{m}_{A\cup B}||^2-\sum_{i\in A}||\vec{x}_i-\vec{m}_A||^2\\ & -\sum_{i\in B}||\vec{x}_i-\vec{m}_B||^2 \\ = & \frac{n_A n_B}{n_A+n_B}||\vec{m}_A-\vec{m}_B||^2,\end{aligned} \tag{1}$$

where A, B denote two clusters in the current iteration, m_A, m_B are the cluster centers, and $\triangle(A,B)$ is the cost of merging clusters A and B.

By default, the hierarchical clustering always groups data points using spatial distance alone, without considering the temporal consistency. This becomes an issue when dealing with manipulation data, which naturally comes with temporal constraints. To alleviate this issue, we apply the Aligned Cluster Analysis (ACA) [30] to reduce the noisiness based on Dynamic Time Alignment Kernel (DTAK) [11], resulting



(a) (b)

Fig. 4: (a) The experimental setup for data collection. We use Vicon system to obtain the poses of human's wrist and object's parts. The camera is used to record the data collection procedure. (b) Visualization of force vectors, which contains both pose and force features.

in a refined segmentation. The ACA is an extension of kernel k-means clustering that could be solved as a versatile energy minimization problem using coordinate descent algorithm:

$$s^* = \operatorname*{argmin}_{s} \mathbf{J}(\mathbf{G}, s) = \sum_{c=1}^{k} \sum_{i=1}^{m} g_{ci} D_c(\mathbf{X}_{[s_i, s_{i+1})}), \tag{2}$$

where $\mathbf{G}_{k \times n}^T \mathbf{1}_k = \mathbf{1}_n$ is the indicator matrix, $g_{ci} = 1$ if sample $\mathbf{X}_i$ belongs to cluster c, and D_c measures the kernel distance between sample point and cluster center. In practice, Equation 2 could be solved in a dynamic programming manner, which leverages the relationship between G and s by solving the Bellman's equation [11]:

$$\mathbf{J}(v) = \min_{v-n_{\max}<i\leq v} (\mathbf{J}(i-1) + \min_{g} \sum_{c=1}^{k} g_c D_\psi^2(\mathbf{X}_{[i,v]}, \dot{\mathbf{z}}_c)), \tag{3}$$

where $D_\psi^2(\mathbf{X}_{[i,v]}, \dot{\mathbf{z}}_c)$ is the squared kernel distance between segment $\mathbf{X}_{i,v}$ and class center c, and $n_{\max}$ defines the maximum segment length of clustering.

B. Event Segmentation

The semantic label of each segment is required to learn a high-level temporal grammar based on the segmented sequences. Although event segmentation of a single segmented motion sequence is straightforward by following its clustering label, it is still difficult to extract the semantic meaning of one segment when having multiple segmented motion sequences performing the same task.

Considering two segmented sequences $\mathbf{X}_{[S_1, S_2 \ldots S_n]}$ and $\mathbf{Y}_{[S_1, S_2 \ldots S_m]}$, we assign semantic labels by merging those segments into clusters where each cluster contains segments that are 'close' in distance. Specifically, we adopt the DTAK [30] criterion $\mathcal{D}(\mathbf{X}_{S_i}, \mathbf{Y}_{S_j})$ to estimate the similarity of segments across different trials of motion primitives segmentation:

$$\mathcal{D}(\mathbf{X}_{S_i}, \mathbf{Y}_{S_j}) = \tau_{\left[\mathbf{X}_{S_i}, \mathbf{Y}_{S_j}\right]}, \tag{4}$$

where $\mathbf{X}_{S_i}, \mathbf{Y}_{S_j}$ are candidate segments that may be grouped together, $\tau_{[\mathbf{X}_{S_i}, \mathbf{Y}_{S_j}]}$ is the similarity metric between two segments calculated recursively using DTAK kernel matrix. Note that it could also be applied to the situation that $\mathbf{X}$ and $\mathbf{Y}$ are the same motion sequence that only differ in segment index i and j.

Based on the distance metric of DTAK, we further apply k-means algorithm to cluster those segments such that each cluster represents one semantic label. The semantic labels of each segmented motion sequence can therefore be obtained by cluster IDs of the corresponding segments.

C. Grammar Induction

After acquiring the semantic labels of multiple segmented motion sequences, we build a T-AOG grammar model using an unsupervised structural learning method [31]. We aim to learn a grammar from a set of sequence of instances that maximize the posterior probability. An initial grammar is built in which the root node is an Or-node, and each branch is an And-node that represents a sequence instance. This initial grammar leads to the maximal likelihood of the training data but has a very small prior probability because of its large size. Starting from the initial grammar, new intermediate non-terminal nodes are generated in a bottom-up fashion to increase its posterior probability. At each iteration, a grammar fragment rooted at a non-terminal node is added into the grammar. In practice, we find it is sufficient to use greedy search with random restarts to identify good grammar fragments.

IV. Inference

Given a sequence of pose and force data Γ as an input, our goal is to find the best motion label sequence A^*, *i.e.*, find the optimal label sequence of the segments that best explains the observation given the learned grammar $\mathcal{G}$ by maximizing the posterior probability:

$$A^* = \operatorname*{argmax}_{A} p(A|\Gamma, \mathcal{G}) = \operatorname*{argmax}_{A} p(\Gamma|A) p(A|\mathcal{G}), \tag{5}$$

where $p(\Gamma|A)$ is the likelihood given the motion label sequence, and $p(A|\mathcal{G})$ is the parsing probability of the parse graph given the grammar. The first term is given by:

$$p(\Gamma|A) = \prod_{k=1}^{K} p(\Gamma_{\gamma_k}|a_{\gamma_k}) = \prod_{k=1}^{K} \prod_{t=t_k^1}^{t_k^2} p(\Gamma_t|a_{\gamma_k}), \tag{6}$$

where k is the segment index, γ_k is the kth segment as introduced in Section II. This term is given by a Gaussian distribution fitted to the learned clusters in the training examples.

The second term $p(A|\mathcal{G})$ in Equation 5 is the Viterbi parsing likelihood, *i.e.*, the probability of the best parse of the string terminals.

Since it is intractable to directly compute the optimal label sequence, we infer the approximately optimal $\widehat{A^*}$ in two steps: i) use the unsupervised clustering method to obtain the segmentation and initialized labels, and ii) refine the labels according to Equation 5 by Gibbs sampling with simulated annealing to find the labeling that maximizes the posterior probability.

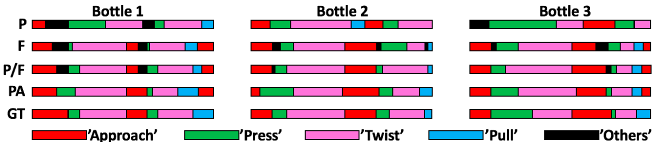


Fig. 5: Qualitative evaluation. Event segmentation and recognition of opening Bottle 1, 2, and 3, from left to right, respectively. P denotes *pose only feature*, F *force only feature*, P/F *force vector feature*, PA *with parsing*, and GT *ground truth*. Each segment represents one type of motion primitive which color is determined by the ground truth sequence.

A. Gibbs Sampling with Simulated Annealing

After initializing the labels by clustering, we find the best parse by Gibbs sampling with simulated annealing. Given an input sequence, we assign one segment label according to the posterior probability (Equation 5) at each iteration. Specifically,

$$a'_{\gamma_k} \sim p(\Gamma_{\gamma_k}|a_{\gamma_k})p(A'|\mathcal{G}), \tag{7}$$

where a'_{γ_k} is the new label of segment Γ_{γ_k}, and A' is the new label sequence obtained by changing the kth label to a'_{γ_k} in the current labeling sequence A. To find the parse with the maximum probability, we adopt simulated annealing to the sampling process by dividing the log probability by a temperature T. We decrease the temperature through the sampling process until the labeling sequence converges.

V. Experiments

A. Human Data Acquisition

a) Tactile Glove: To capture both pose and force in hand-object interactions, we utilize an open-source tactile glove [3]. The tactile glove employs a network of 15 IMUs to measure the rotations between individual phalanxes. Hand pose is reconstructed using forward kinematics. With 6 customized force sensors using Velostat, a piezoresistive material, the force exerted by hand is recorded in two regions (proximal and distal) on each phalange and a 4×4 regions on the palm. The data is collected and visualized using the Robot Operating System (ROS).

b) Experimental Setup: We utilize a Vicon motion capture system to obtain the relative poses between the wrist of hand and object parts. Fig. 4a describes the schematic of the experimental environment setup in human data acquisition. Six Vicon cameras are placed on top left and top right in front of the area of interests.

c) Force Vectors: Force vectors are computed as the extracted features from the force and pose data (see Fig. 4b). Each force scalar measured on hand is normalized and treated as the magnitude of the force vector. The orientation of the force vector is set to be perpendicular to the fingers. All the force vectors are expressed with respect to one fixed frame by applying the chain product of homogeneous transforms. Hence, we are able to combine the heterogeneous pose and force information into one compact form of feature vector.

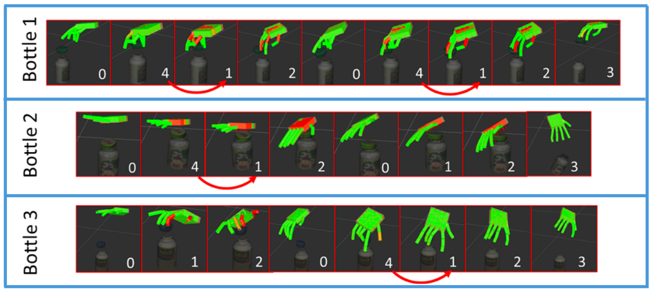


Fig. 6: Key frames of opening various bottles with T-AOG. The numbers indicate the cluster labels and the red arrows indicate the merges triggered by the parsing of T-AOG.

B. Evaluation

The performance is evaluated by the frame-wise recognition accuracy, *i.e.*, comparing the predicted event label with the ground truth frame by frame. The ground truth segmentation is manually labeled. Based on this protocol, we evaluate the correspondence in three metrics: i) *Pose* feature as the Euler angles of each phalanx, ii) *Force* feature as the magnitude of the force, and iii) the combination of *Pose* and *Force* in the form of force vectors. For fair comparison, the results reported below use the cluster number $k=5$ and maximum segment length $n_{max}=200$.

C. Event Segmentation and Recognition with Clustering

Fig. 5 visualizes the event recognition results by segmenting each motion primitive of the trials in opening *Bottle 1, 2,* and *3*. Quantitative results are shown in Table I. The segmentation using only pose data has the worst performance compared with the ground truth. The use of force data shows a significant improvement compared to the pose only data. This result indicates the benefits of the force information during hand-object manipulation. Combining both pose and force data together outperforms that only uses either pose or force data.

D. Segmentation, Recognition and Parsing with T-AOG

To further reduce noise, mislabeling, and incoherence, T-AOG is integrated to refine the segmentation, recognition and parsing of the motion sequences by maximizing the posterior probability.

TABLE I: Quantitative Evaluation. With clustering only, we use the hand pose, in the forms of Euler angles of each phalanx; hand force, as scalars; and the combination of pose and force as force vectors as feature inputs. Including force factor yields higher correspondence with ground truth sequence. Parsing the events with T-AOG on top of the clustering, the performance improves significantly.

	Clustering only			With T-AOG
	Pose only	Force only	Pose and Force	Pose and Force
Bottle 1	55.3%	67.5%	70.3%	**78.6%**
Bottle 2	62.0%	70.9%	76.2%	**82.5%**
Bottle 3	54.1%	71.1%	72.9%	**78.5%**

Fig. 6 shows the motion frames during the interactions in opening three types of bottles. The number in each frame denotes its motion label which is produced by the proposed clustering pipeline. Additionally, we highlight the changes after applying the proposed annealing inference framework, indicated by the red arrow, which reveals the directions of label merging.

Experimental results after integrating a T-AOG parsing are both qualitatively and quantitatively presented. As depicted in Fig. 5, comparing to model-free clustering methods, the T-AOG based parsing approach recovers some noisy and mislabeled segments, resulting in more coherent results. Last column of Table I shows the quantitative results. The performance of both segmentation and recognition using T-AOG have a marked improvement compared to the methods only by clustering, demonstrating the usefulness of learning a grammar model for events parsing and inference.

VI. Conclusion and Future Work

In this paper, we present an unsupervised approach for manipulation event segmentation, recognition and parsing. Hand-object interaction sequences are segmented in an unsupervised learning fashion, based on which a temporal grammar is further induced. Through a tactile glove, our work explicitly incorporates forces imposed by hands in addition to its pose.

The experiments demonstrate that force is indeed an important factor as it significantly improves motion primitives segmentation. In addition, learning a grammar model T-AOG from the clustering results for parsing the motions can reduce noisiness and eliminate mislabeling and ultimately lead to a more coherent event segmentation and parsing.

In the future, the proposed approach could be used to improve the traditional event segmentation, recognition and parsing in computer vision by inferring the force from the videos [8]. It is also possible to use the segmentation as the demonstrations to teach robots with LfD to open medicine bottles [32] or more complex tasks, *e.g.*, tool uses [33] and folding clothes [23].

Acknowledgement: The work reported herein was supported by DARPA XAI grant N66001-17-2-4029 and ONR MURI grant N00014-16-1-2007.

References

[1] R. Poppe, “A survey on vision-based human action recognition,” *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[2] D. Weinland, R. Ronfard, and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[3] 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*, 2017.

[4] 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.

[5] 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.

[6] Y. Zhu, Y. Zhao, and S. Chun Zhu, “Understanding tools: Task-oriented object modeling, learning and recognition,” in *CVPR*, 2015.

[7] 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.

[8] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S.-C. Zhu, “Inferring forces and learning human utilities from videos,” in *CVPR*, 2016.

[9] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, “Real-time human pose recognition in parts from single depth images,” *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[10] P. Wei, N. Zheng, Y. Zhao, and S.-C. Zhu, “Concurrent action detection with structural prediction,” in *ICCV*, 2013.

[11] F. Zhou, F. De la Torre, and J. K. Hodgins, “Hierarchical aligned cluster analysis for temporal clustering of human motion,” *PAMI*, vol. 35, no. 3, pp. 582–596, 2013.

[12] J. Wang, Z. Liu, and Y. Wu, “Learning actionlet ensemble for 3d human action recognition,” in *Human Action Recognition with Depth Cameras*, pp. 11–40, Springer, 2014.

[13] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, “Cross-view action modeling, learning and recognition,” in *CVPR*, 2014.

[14] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, “Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization,” *PAMI*, vol. 39, no. 6, pp. 1165–1179, 2017.

[15] S. Qi, S. Huang, P. Wei, and S.-C. Zhu, “Predicting human activities using stochastic grammar,” in *ICCV*, 2017.

[16] M. A. Brubaker and D. J. Fleet, “The kneed walker for human pose tracking,” in *CVPR*, 2008.

[17] M. A. Brubaker, L. Sigal, and D. J. Fleet, “Estimating contact dynamics,” in *ICCV*, 2009.

[18] M. A. Brubaker, D. J. Fleet, and A. Hertzmann, “Physics-based person tracking using the anthropomorphic walker,” *IJCV*, vol. 87, no. 1, pp. 140–155, 2010.

[19] 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.

[20] J. Bandera, J. Rodriguez, L. Molina-Tanco, and A. Bandera, “A survey of vision-based architectures for robot learning by imitation,” *International Journal of Humanoid Robotics*, vol. 9, no. 01, p. 1250006, 2012.

[21] Y. Yang, Y. Li, C. Fermüller, and Y. Aloimonos, “Robot learning manipulation action plans by” watching” unconstrained videos from the world wide web.,” in *AAAI*, 2015.

[22] L. Pinto, D. Gandhi, Y. Han, Y.-L. Park, and A. Gupta, “The curious robot: Learning visual representations via physical interactions,” in *ECCV*, 2016.

[23] C. Xiong, N. Shukla, W. Xiong, and S.-C. Zhu, “Robot learning with a spatial, temporal, and causal and-or graph,” in *ICRA*, 2016.

[24] S. Manschitz, J. Kober, M. Gienger, and J. Peters, “Learning to sequence movement primitives from demonstrations,” in *IROS*, 2014.

[25] S. Manschitz, M. Gienger, J. Kober, and J. Peters, “Probabilistic decomposition of sequential force interaction tasks into movement primitives,” in *IROS*, IEEE, 2016.

[26] Y. Chebotar, O. Kroemer, and J. Peters, “Learning robot tactile sensing for object manipulation,” in *IROS*, 2014.

[27] A. X. Lee, H. Lu, A. Gupta, S. Levine, and P. Abbeel, “Learning force-based manipulation of deformable objects from multiple demonstrations,” in *ICRA*, 2015.

[28] L. Rozo, S. Calinon, D. G. Caldwell, P. Jimenez, and C. Torras, “Learning physical collaborative robot behaviors from human demonstrations,” *IEEE Transactions on Robotics*, vol. 32, no. 3, pp. 513–527, 2016.

[29] 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.

[30] F. Zhou, F. De la Torre, and J. K. Hodgins, “Aligned cluster analysis for temporal segmentation of human motion,” in *Automatic Face & Gesture Recognition, 2008. FG’08. 8th IEEE International Conference on*, 2008.

[31] K. Tu, M. Pavlovskaia, and S.-C. Zhu, “Unsupervised structure learning of stochastic and-or grammars,” in *NIPS*, 2013.

[32] M. Edmonds, F. Gao, X. Xie, H. Liu, S. Qi, Y. Zhu, B. Rothrock, and S.-C. Zhu, “Seeing the force: Integrating poses and visually latent forces for learning manipulations through fluent discovery,” in *IROS*, 2017.

[33] Y. Zhu, Y. Zhao, and S.-C. Zhu, “Understanding tools: Task-oriented object modeling, learning and recognition,” in *CVPR*, 2015.