Unsupervised Learning of Hierarchical Models for Hand-Object Interactions

[1]Center for Vision, Cognition, Learning, and Autonomy - University of California, Los Angeles

[2]Jet Propulsion Laboratory, California Institute of Technology, Los Angeles

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]

Introduction

We present an unsupervised learning method for manipulation event segmentation, recognition and parsing. By using a self-made tactile glove, we can reliably retrieve contact force during hand-object manipulations.

The proposed method is able to:

- Incorporate invisible force of hand manipulation for event segmentation and parsing.
- Unsupervisedly learn a temporal grammar model (T-AOG) for motion recognition.
- Model noisy and heterogeneous hand sensory data.

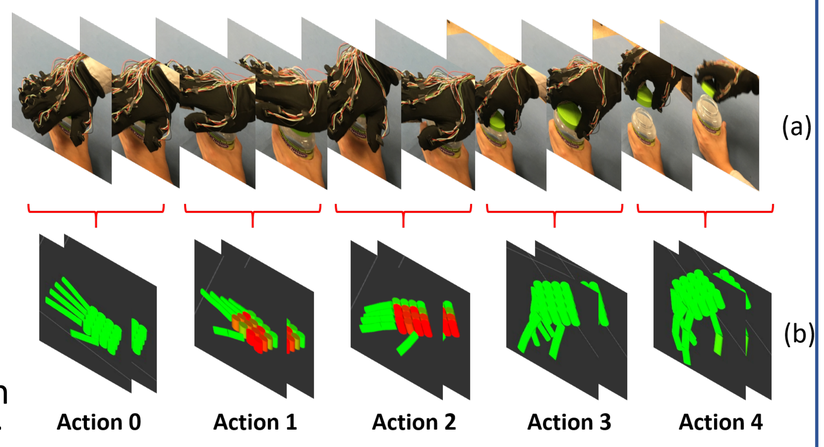

Fig 1. (a) A sequence of movement primitive demonstrated by an agent for a manipulation task – opening a medicine bottle captured by a tactile glove. (b) Reconstructed force and pose data using the tactile glove. Our proposed method segments and parses the noisy inputs of force and pose in an unsupervised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

Representation

We introduce a structural grammar model Temporal And-Or graph (T-AOG) to represent the temporal structure of a task. And-node is decomposed into sub-events or motion primitives as its child nodes. Or-node encodes alternative solutions to perform a sub-task. A pg is a sub-graph of T-AOG that captures the temporal structure of the scenario.

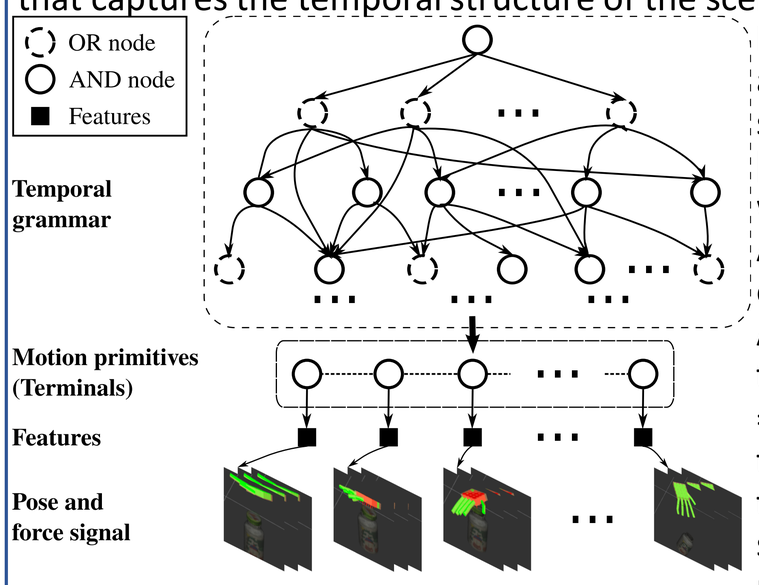

Fig 2. Illustration of the T-AOG. The terminals are motion primitives of hand-object interactions.

Pose and Force features Γ are extracted based on a raw sensory sequence $\mathbf{I}$ in time [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 sequence is defined as $\boldsymbol{\Gamma}=\{\boldsymbol{\gamma}_k\}$, k=1,…,K. $\boldsymbol{\gamma}_k=[t_k^1, t_k^1]$ is the time interval in which the motion primitive are the same. $a_{\gamma k}$ denotes the motion label for the segment $\mathbf{I}_{\gamma k}$.

Unsupervised Learning of Hand-Object Motion Primitives

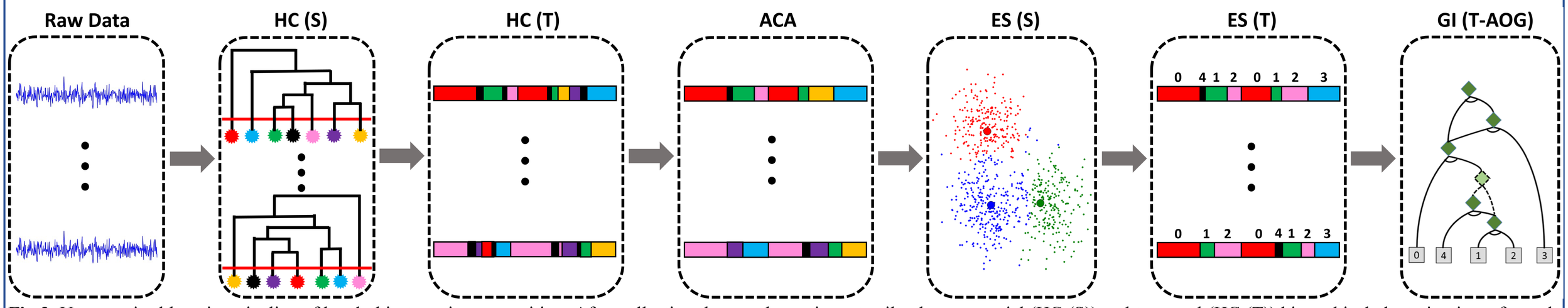

Fig 3. Unsupervised learning pipeline of hand-object motion recognition. After collecting the raw data using a tactile glove, a spatial (HC (S)) and temporal (HC (T)) hierarchical clustering is performed on both force and pose data. An aligned cluster analysis (ACA) is adopted to further reduce the noise. 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.

The pipeline starts from Hierarchical Clustering where we adopt Wards method to determine clusters merging. Considering temporal consistency of clustered segments, Aligned Clustering Analysis is applied based on Dynamic Time Alignment Kernel (DTAK). It solves the kernel k-means clustering as a versatile energy minimization problem using coordinate descent algorithm. To generate semantic label of each segment, we estimate DTAK similarity of segments across different trials of motion primitives segmentation. Then T-AOG grammar model is built on those motion sequences with semantic labels.

Inference

Given a sequence of force and pose data Γ as input, our goal is to find the optimal motion label sequence A^* that best explains the observation based on learned grammar $\boldsymbol{g}$ by maximizing the posterior probability:

$$A^* = arg\max_A p(\Gamma|A)p(A|g), \quad (1)$$

$$P(\Gamma|A) = \prod_{k=1}^{K} p(\Gamma_{\gamma k}|a_{\gamma k}) = \prod_{t=t_k^1}^{t_k^2} p(\Gamma_t|a_{\gamma k}), \quad (2)$$

$P(\Gamma|A)$ is the likelihood given the motion label sequence, $p(A|g)$ is the parsing probability of the parse graph given the grammar.

Fig 4. Key frames of opening various bottles with T-AOG. Numbers on bottom right indicate the cluster labels and the red arrows indicate the merges triggered by the parsing of T-AOG.

We infer the optimal A^* in two steps: i) use clustering method to obtain the segmentation and initialized labels, and ii) refine the labels according to Eq. (1) by Gibbs sampling with simulated annealing that maximizes the posterior probability.

Evaluation

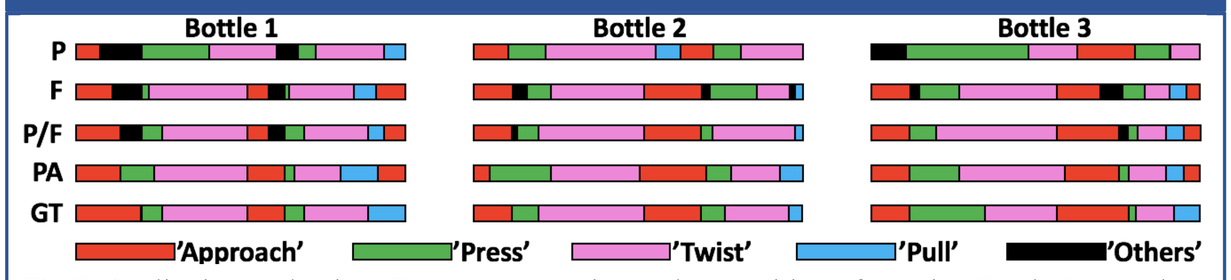

Fig 5. Qualitative evaluation. Event segmentation and recognition of opening Bottle 1, 2, and 3, from left to right, respectively. P denotes *pose only feature*, F *force only feature*, P/F *force vector feature*, PA *with parsing*, and GT *ground truth*. Each segment represents one type of motion primitive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

	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%**

TABLE 1. Quantitative evaluation. With clustering only, we use 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 with T-AOG on top of clustering, the performance improves significantly.

The performance is evaluated by the frame-wise recognition accuracy. The ground truth segmentation is manually labeled in ROS RVIZ. The results reported use the same cluster number *K=5* for fair comparison.