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.

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.

Pose and force features $\Gamma$ are extracted based on a raw sensory sequence $I$ in time $[i, t]$. Each frame is labeled with motion primitive $a_k$. Aggregating together, we obtain a label sequence $A=(a_k)$. The segmentation of the sequence is defined as $I_k=[y_1, ... , y_m]$, $y_m=[I_{t_k}, I_{t_{k+1}}]$ is the time interval in which the motion primitive are the same. $a_{k_i}$ denotes the motion label for the segment $I_{P_k}$.

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.