USC Viterbi School of Engineering | Ming Hsieh Department of Electrical Engineering

Hand Graph Representations for Unsupervised Segmentation of Complex Activities

Pratyusha Das[1], Jiun-Yu Kao[1], Antonio Ortega[1], Tomoya Sawada[2], Hassan Mansour[3], Anthony Vetro[3], Akira Minezawa[2]

[1] University of Southern California (USA), [2] Mitsubishi Electric Corporation (Japan), [3] Mitsubishi Electric Research Labs (USA)

Introduction

- **Motivation:**
  - Significant development in generic video based human motion tracking with Openpose allows us to use this as preprocessing tool to extract 2D hand keypoints from the video.
  - It also takes care of the privacies of the scene.
- **Contribution:**
  - Graph representation of hand skeleton data is introduced.
  - A new fine complex motor activity hand dataset of an assembling task is introduced and made public for research community.
  - Unsupervised temporal segmentation of a sequence of complex sub-tasks using is proposed in order to evaluate the efficiency of an assembly task.

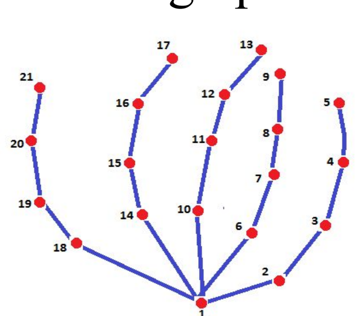
Proposed hand graph features

- Symmetric graph laplacian $\boldsymbol{L}$ is define as $L = I - D^{-1/2}AD^{-1/2}$ where $\boldsymbol{A}$ and $\boldsymbol{D}$ represent adjacency matrix, degree matrix respectively.
- Spectral basis of the graph $u_1, u_2, \ldots, u_{N_v}$: Eigen vectors of $\boldsymbol{L}$, leading to the columns of matrix $\boldsymbol{U}$.
- Spectral frequencies [3] are the corresponding eigen values.
- Graph signal is represented as linear combination of u_k .

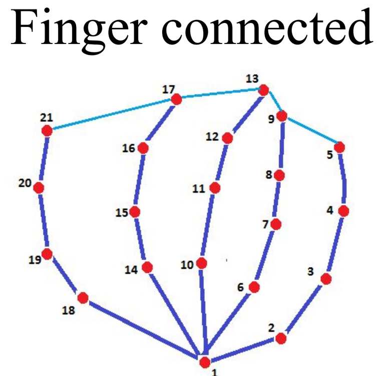
$$c_i = \sum_{k=1}^{N_v} \alpha_{k,i}\, u_k \quad \text{and} \quad \alpha_{k,i} = c_i^T u_k$$

- c_i : the motion vector present in each node of hand, $\alpha_{k,i}$: graph fourier coefficients, a unique representation of the motion vectors, used as graph features.

$\mathcal{G}_{\mathcal{H}}$: Hand graph



$\mathcal{G}_{\mathcal{FH}}$: Finger connected hand graph



$\mathcal{G}_{\mathcal{LRH}}$: Left-Right hand graph

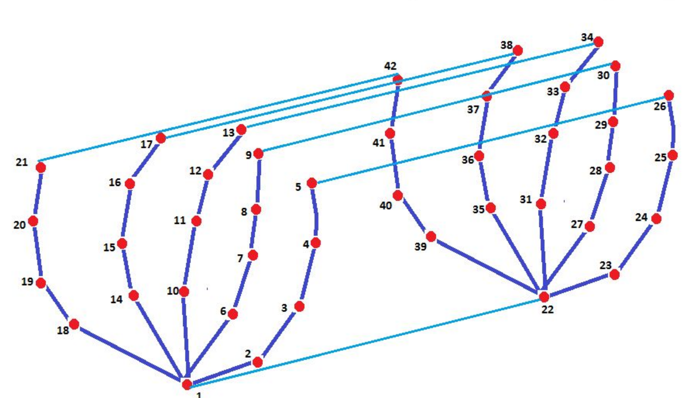


Fig 1. Proposed hand graphs (Fixed, undirected, unweighted)

- $\mathcal{G}_{\mathcal{FH}}$ is constructed in order to account for relative motion of the tips of the fingers.
- $\mathcal{G}_{\mathcal{LRH}}$ can capture the relative motion between two hands along with the intra-hand motion.

Unsupervised online segmentation

- Bayesian Information criterion (BIC) [2] based unsupervised online segmentation algorithm is used.
- At time point i, Generalized likelihood ratio (GLR) between feature matrix of left (W_l) and right (W_r) window of i is computed.

$$\Delta BIC_i = \log\left(\frac{|\boldsymbol{\Sigma}_{\boldsymbol{W}_l \cup \boldsymbol{W}_r}|^{\frac{N}{2}}}{|\boldsymbol{\Sigma}_{\boldsymbol{W}_l}|^{\frac{N_l}{2}}\,|\boldsymbol{\Sigma}_{\boldsymbol{W}_r}|^{\frac{N_r}{2}}}\right) - \frac{\lambda}{2}\left(d + \frac{d(d+1)}{2}\right)\log N$$

Here, Σ is the covariance matrix, d is the feature dimension, N is the length of the data sequence, and λ controls the number of segments.

- $\Delta BIC_i \lessgtr 0$ decides i is a good segmentation instant or not.
- If i is not a segmentation instant, we combine W_l and W_r , and go to $i+1$ to check with the next window.

Experimental setup and dataset

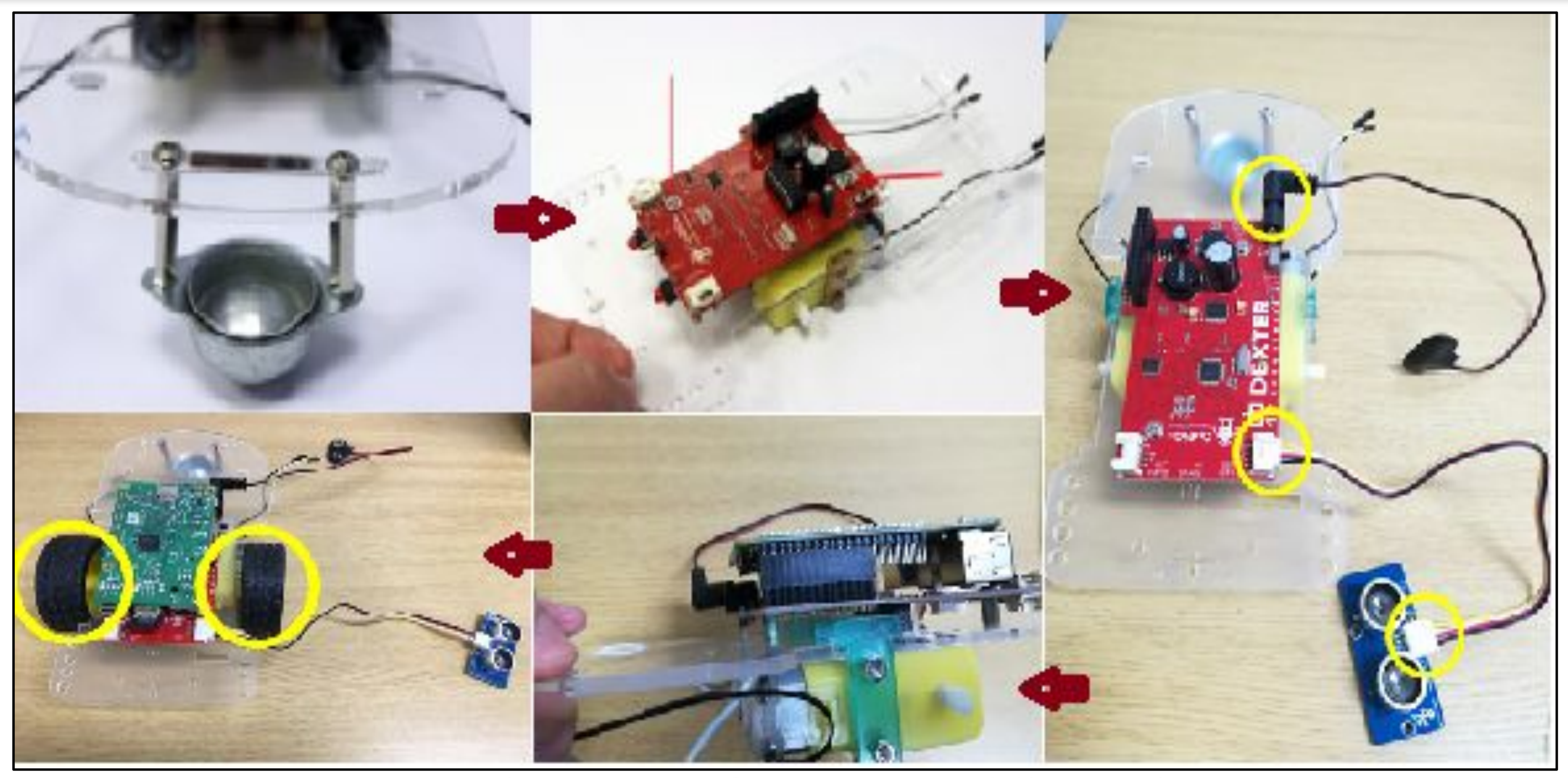


Fig. 2 Steps for toy assembling task

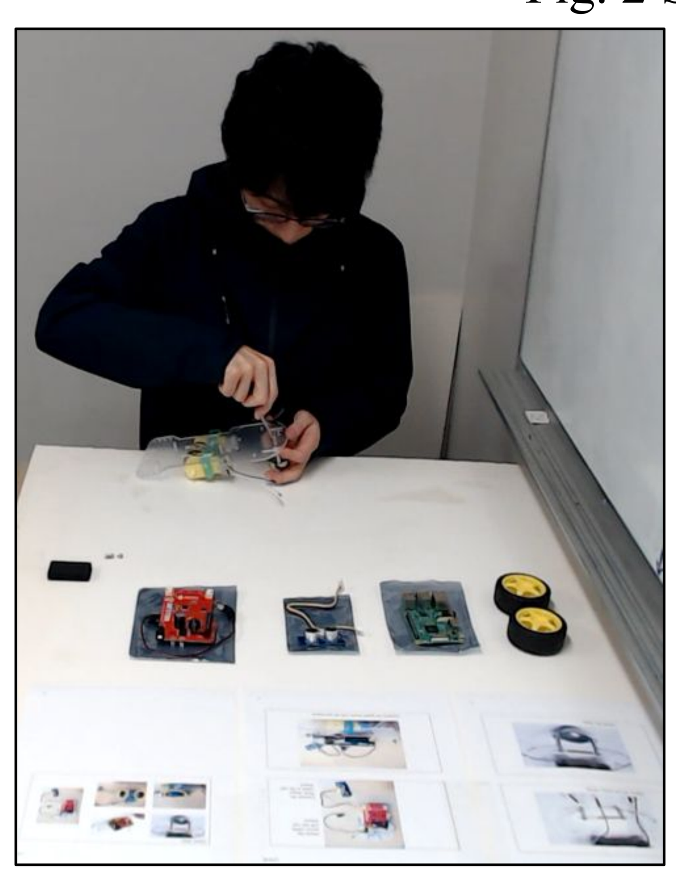


Fig 3. Participant performing the task

- A toy assembling task with three subtasks: Assembling (involves use of screws), Combining (involves use of wires and pins) and Checking.
- No. of participants : 11, each performing the task 3 times.
- Total no. of data sequences : 33.
- Openpose, used as a preprocessing tool to extract 2D position of 2X21 hand keypoints from the video at fps 30.
- 2D motion vectors from the position data is computed as it captures all the fundamental variation present in each sub-tasks.

Results

- True segmentation instances $\hat{S}_{a_i}$ are the segmentation instances which lie in a segmentation zone around the ground truth segmentation points S_{g_i}.
- Segmentation accuracy is defined as the number of frames which are grouped correctly to the total number of frames.
- To take into consideration the early and late segmentation, S_1 is computed using following equation, where L is the length of the sequence.

$$S_1 = \left(1 - \sum_{i=1}^{L_g} \beta_i \frac{S_{g_i} - \hat{S}_{a_i}}{L}\right) X100$$

- In this scenario, as only 2D position data of hand keypoints is accessible, we have very limited information about the scene, thus using the motion vectors as features in baseline evaluation.
- Graph $\mathcal{G}_{\mathcal{LRH}}$ outperformed other graphs and baseline method.

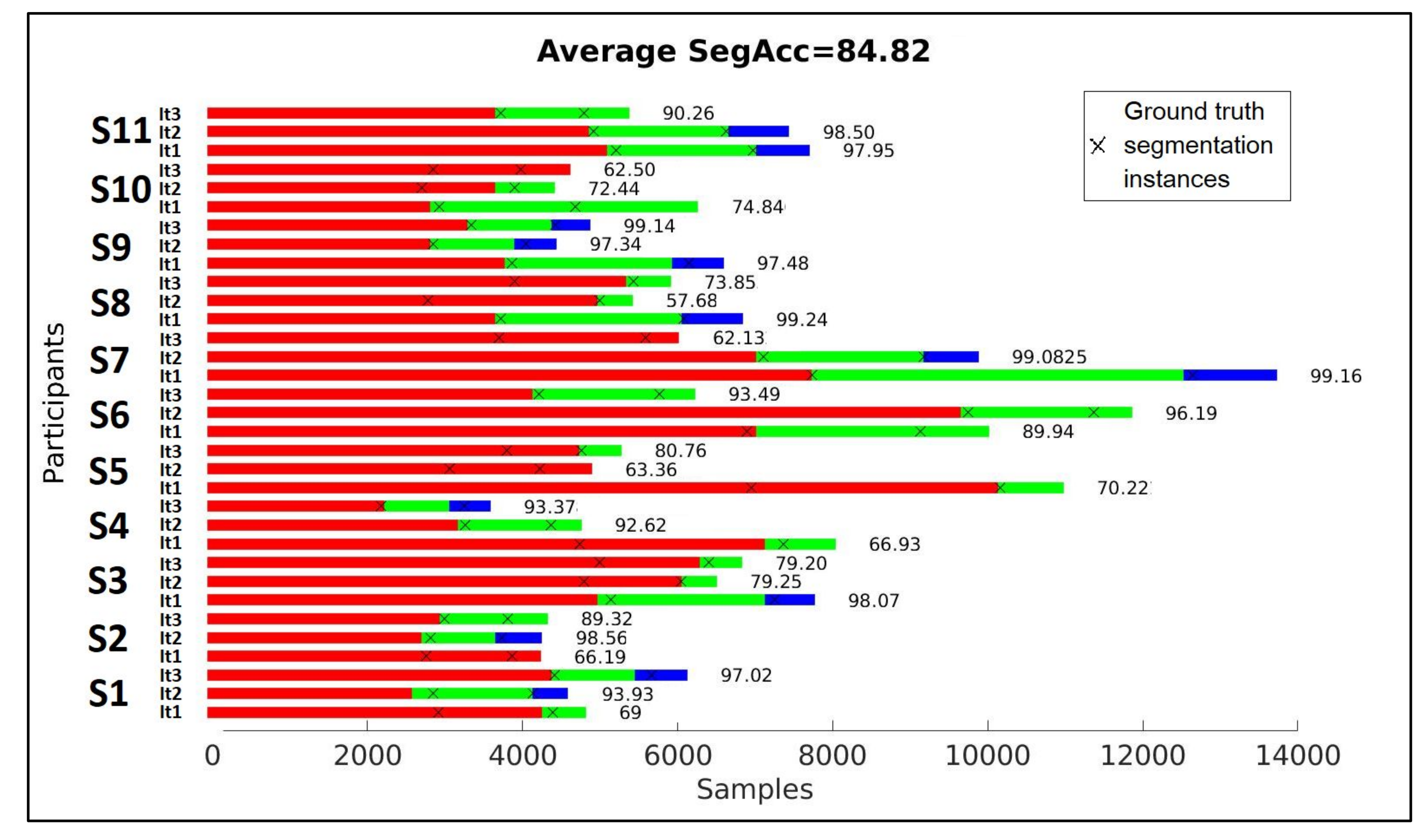


Fig 4. Segmentation outcome using features from graphs, transition in color represents change in action

Table 1. Summarized results (%)

Method	Precision	Recall	F1-score	SegAcc	S_1
Baseline	25.1	33.3	22.2	71.58	16.4
Proposed	**54.3**	**85.7**	**64.1**	**84.8**	**59.6**

Future work

- Qualitative analysis of the performance of the participants in the context of segmentation.
- Explore the choice of weighted hand graphs.

[1] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
[2] Han, Kyu J., Panayiotis G. Georgiou, and Shrikanth S. Narayanan. "The SAIL speaker diarization system for analysis of spontaneous meetings." *2008 IEEE 10th Workshop on Multimedia Signal Processing*. IEEE, 2008.
[3] Kao, Jiun-Yu, Antonio Ortega, and Shrikanth S. Narayanan. "Graph-based approach for motion capture data representation and analysis." *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014..

ICASSP 2019