This essay introduces a novel algorithm that could segment and recognize surgical operation trajectories into distinct, meaningful gestures. Segmenting a task into gestures can provide detailed feedback, for example when and where unexpected cuts and minor injuries happened, not just an overall ‘successor, failure’ feedback, so this step is preliminary and critical in surgical workflow analysis. The result can be used for facilitating learning from demonstrations for autonomous robotic surgery and other robot-assist surgical tasks.

In this paper, we propose a hierarchical semi-supervised learning framework for surgical gesture segmentation. In the first hierarchy, we identify the critical points from kinematics data and video data and then determine the potential segmentation points. In the second hierarchy, we use an unsupervised learning approach to cluster the potential segmentation points and determine the final segmentation points.

Our method is tested on JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS)

[18], where a large number of surgical operation data was collected from surgeons using the da Vinci Surgical Robot, containing kinematics and video data along with manual annotation with gestures’ name

After filtering noise data, we calculate Euclidean distance for translation data and quaternion angular distance for rotation data. As shown in Figure with trajectory and ground truth segment points, segmentation points tend to cluster, while points within continuous motion are typically sparse [20]. These characteristics are frame invariant, indicating that they remain consistent when the data is transformed into a new frame of reference. Hence, calculating variances across all new frames can help identify significant changes.

A modified Transformer-based architecture integrated with ResNet-18 backbones with pre-trained weights is used to implement visual feature extraction in this paper. The remaining weights in the architecture are fine-tuned using the limited labeled data in JIGSAWS.

The feature clustering model is a two layers of DP-GMM model. The first layer cluster across all the frames to find change points as much as possible, while the second layer further clusters those change points to find final segmentation points.

Combines the output of both kinematic and vision potential segmentation points, we use another DBSCAN to identify the final segmentation points.

Bayesian optimization is used to automatically find the optimized values of hyperparameters [23]. Specifically, the concentration parameters and component numbers for the two DP-GMM models, and the maximum distance for DBSCAN can be fine-tuned through Bayesian optimization.

The dataset consists a total of 103 demonstration files, which are divided into a training set, a validation set, and a test set in a roughly 7:2:1 ratio. Compared with State-of-the-Art Methods, our proposed method achieving an impressive accuracy of 0.856 for classification and an F1 score of 0.623 for segmentation. outperforms the baseline techniques in terms of accuracy. In the future, we plan to expand the algorithm by incorporating self-supervised methods and leveraging sim-to-real learning techniques to further eliminate the need for labeling real operation data for surgical gesture recognition tasks.