Hierarchical Semi-Supervised Learning Framework for Surgical Gesture Segmentation and Recognition Based on Multi-Modality Data

Zhili Yuan, Jialin Lin, Dandan Zhang

Abstract— **Segmenting and recognizing surgical operation trajectories into distinct, meaningful gestures is a critical preliminary step in surgical workflow analysis for robot-assisted surgery. This step is necessary for facilitating learning from demonstrations for autonomous robotic surgery, evaluating surgical skills, and so on. In this work, we develop a hierarchical semi-supervised learning framework for surgical gesture segmentation using multi-modality data (i.e. kinematics and vision data). More specifically, surgical tasks are initially segmented based on distance characteristics-based profiles and variance characteristics-based profiles constructed using kinematics data. Subsequently, a Transformer-based network with a pre-trained 'ResNet-18' backbone is used to extract visual features from the surgical operation videos. By combining the potential segmentation points obtained from both modalities, we can determine the final segmentation points. Furthermore, gesture recognition can be implemented based on supervised learning.**

The proposed approach has been evaluated using data from the publicly available JIGSAWS database, including Suturing, Needle Passing, and Knot Tying tasks. The results reveal an average F1 score of 0.623 for segmentation and an accuracy of 0.856 for recognition. For more details about this paper, please visit our website: `https://sites.google.com/view/surseg/home`.

I. Introduction

The rapid development of machine learning leads to substantial growth in the field of surgical data science and robot-assisted surgery [1]–[4]. Surgical tasks mainly rely on the repetition and execution of specific gestures, while they can be further decomposed into basic surgical gestures. The segmentation and recognition of surgical gestures is an important task for surgical workflow analysis, which can benefit the training and assessment of surgeons [5], enable semi-autonomous robotic surgery via learning from demonstration, and provide real-time feedback and guidance for surgeons to enhance the efficiency of robotic surgery [6].

Supervised learning has been widely used for surgical gesture segmentation and recognition [7]. Traditional machine learning algorithms, such as Hidden Markov Models (HMM) [8], Linear Dynamical Systems (LSD), Conditional Random Fields (CRF), and Markov/semi-Markov CRF (MsM-CRF), have been used for surgical gesture segmentation and recognition [9]. Recurrent Neural Networks (RNNs) and their variations [10], such as Long-Short Term Memory (LSTM), and Gated Recurrent Units (GRU), have also been widely used for gesture segmentation, since they have been proven to be effective for processing sequential data [11]. However, the machine learning-based gesture segmentation quality significantly depends on the availability of large-scale labeled datasets.

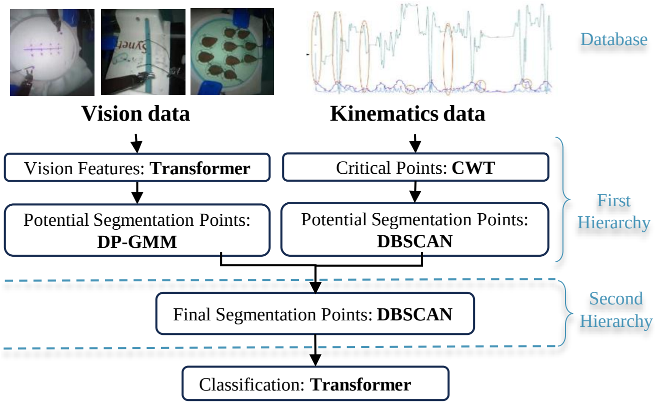
Vision data
Kinematics data
Database
Vision Features: Transformer
Critical Points: CWT
Potential Segmentation Points: DP-GMM
Potential Segmentation Points: DBSCAN
First Hierarchy
Final Segmentation Points: DBSCAN
Second Hierarchy
Classification: Transformer

Fig. 1. Flowchart of the proposed hierarchical semi-supervised learning framework for surgical gesture segmentation and recognition.

Unsupervised learning has been explored for surgical gesture segmentation. For example, Transition State Clustering (TSC) has been applied to the surgical gesture segmentation task [12]. Gaussian Mixture Model has been used to segment surgical gestures based on kinematic data and the visual features extracted by pre-trained Convolutional Neural Networks (CNN) [13]. A Dense Convolutional Encoder-Decoder Network (DCED-Net) has been combined with Temporal Convolutional Networks (TCN) to further enhance the segmentation accuracy [14]. However, the performance of unsupervised learning methods is not desirable compared to supervised learning methods. As a compromise between data labeling workload and model performance, we investigate semi-supervised learning in this paper.

Semi-supervised learning has been applied to the segmentation of surgical gestures using kinematic data. For instance, within a semi-supervised learning framework, the Stacked Denoising Autoencoder (SDAE) has been utilized for feature extraction in an unsupervised manner [15], while Dynamic Time Warping (DTW) was employed to align the kinematic data from different trials of the surgical task. Subsequently, a voting mechanism based on kernel density estimation was applied to transfer labels from template trials to the test trial, resulting in gesture recognition through semi-supervised learning. An RNN-based generative model has been developed for surgical gesture recognition, using only one annotated sequence [16]. Its performance outperforms other approaches, including the RNN-Based Autoencoder and RNN-Based Future Prediction. However, most of the methods mentioned above only use kinematic data to imple-

All authors are with the Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom.

ment gesture segmentation or recognition. Recent literature has demonstrated that the performance of surgical gesture recognition can be improved using multi-modality data compared with their single-modality counterparts [13].

To this end, we propose a hierarchical semi-supervised learning framework for surgical gesture segmentation and recognition using multi-modality data, which aims to eliminate the need of collecting a large amount of labeled data for supervised learning while ensuring high segmentation and recognition performance. The main contributions are listed as follows.

- We propose a hierarchical semi-supervised learning framework for surgical gesture segmentation, utilizing both kinematics and video data. Our proposed method was compared to state-of-the-art approaches, with results indicating higher accuracy of segmentation.
- We employ a transfer learning approach to extract useful features from a limited amount of labeled surgical video data, thus ensuring high data efficiency in our proposed method. This method utilizes a Transformer-based architecture, with ResNet-18 serving as the backbone.

II. Methodology

A. Overview

The core target of surgical gesture segmentation is to locate the starting point and ending point of a specific surgical gesture [17]. Surgical gestures involve dynamic movements that feature transitions encompassing both local and global motions. As such, kinematic data play a crucial role in detecting transitional states and is thus considered valuable for surgical gesture segmentation. Simultaneously, vision data, with contextual information, contributes to improving the accuracy of surgical activity segmentation and recognition. Therefore, multi-modality data will be used to enhance surgical gesture segmentation and recognition accuracy in this paper.

We define features initially identified from specific feature extraction methods as **‘critical points’** (also known as ‘change points’ or ‘transition points’ in other papers). It is worth noting that the occurrence of critical points may not be simultaneous for different characteristics profiles. For instance, in a Knot Tying task, orienting a needle may only lead to changes in the rotation, while the translation may remain unchanged. As a result, we introduce a hierarchical structure to address this issue. In the first layer, multiple meaningful features extracted from kinematic data are clustered using ‘Density-Based Spatial Clustering of Applications with Noise (DBSCAN)’, while all visual features extracted by the Transformer-based model are clustered with Dirichlet Process Gaussian Mixture Model (DP-GMM). The resulting data will be referred to as **‘potential segmentation points’** (also known as ‘pre-segmentation points’ in other papers) throughout this paper. In the final step, the second layer of DBSCAN is applied to all potential segmentation points to generate **‘final segmentation points’**. **Algorithm 1** describes the whole segmentation method.

Algorithm 1: Overview of the Proposed Framework

```
for All kinematics data do
    Obtain Filtered Kinematics Data:
    Raw Data ← Kalman and Savitzky-Golay filters
    Obtain critical points $cp^{ori}$: Distance
    Characteristics-Based Profiles ← CWT
    Obtain critical points $cp^{var}$: Variance
    Characteristics-Based Profiles ← CWT
    (see Section II-B)
end for
for All vision data do
    Extract visual features $f_{vis}$:
    Preprocessed video data ← Transformer
    (see Section II-C.1)
    Obtain all potential segmentation points for vision
    data $CP_{vis}$:
    $f_{vis}$ ← DP-GMM (see Section II-C.2)
end for
Obtain all potential segmentation points for kinematics
data $CP_{kine}$:
$[cp^{ori}, cp^{var}]$ ← DBSCAN $1_{st}$ Layer
Obtain final segmentation points $esp$:
$[CP_{kine}, CP_{vis}]$ ← DBSCAN $2_{nd}$ Layer
(See Section II-D)
```

One of the well-known surgical activity datasets is the 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. This dataset contains kinematics and video data along with manual annotation with gestures’ name and operators’ skill levels. There are three surgical tasks in the dataset, including 39 suturing, 36 knot tying and 20 needle passing demonstrations. The definitions of the surgical gesture for all three tasks can be found in [18]. The JIGSAWS dataset is used for model training and evaluation in this paper.

B. *Pre-Segmentation Based on Kinematics Data*

1) *Characteristics Profile Construction:*

The kinematic characteristics of the surgical gesture transition points can differ from other points within the same gesture [19]. That is to say, kinematic data points are clustered when they are transitioning to another gesture, while smooth intervals represent an in-motion state. By calculating the translation and rotation distance between two consecutive points along the trajectory, we can find out critical points (transition points) in the density of trajectories. In addition, a variance characteristics-based method is adopted to enable the segmentation to be more robust to noise [20]. The raw kinematics data recorded include noises. Therefore, the Kalman filter is applied to the raw trajectories for noise removal. Following that, new profiles can be constructed from distance and variance characteristics-based profiles. The Savitzky–Golay filter can then be applied to further smooth the newly constructed profiles after normalization.

Denote the original frame for the kinematics data as B, Translation Matrix as $\mathbf{T(t)}$, and 3×3 Rotation Matrix as $\mathbf{R}(t)$. Suppose that $\mathbf{Q(t)}$ is the quaternion converted from $\mathbf{R}(t)$, which also represents the end-effector's orientation in the original dataset. The corresponding Euler angles converted by $\mathbf{R}(t)$ is $\mathbf{e}(t) = [e_x(t), e_y(t), e_z(t)]$. Let $\mathbf{p}(t) = [p_x(t), p_y(t), p_z(t)]$ denote the end-effector position at time step t, $\dot{\mathbf{p}}(t) = [\dot{p_x}(t), \dot{p_y}(t), \dot{p_z}(t)]$ denote the end-effector velocity at time step t.

Distance Characteristics-Based Profiles: Translation and rotation distance between consequent time steps can be calculated for both left and right surgical tools, respectively. For the translation component, the Euclidean distance of the end-effector between time step t and $t-1$ can be calculated by $D_{trans} = ||\mathbf{p}(t) - \mathbf{p}(t-1)||$. As for the rotational component, the distance is determined by $D_{rot} = arccos(2(\mathbf{Q}(t) \cdot \mathbf{Q}(t-1))^2 - 1)$.

Four distance characteristics-based profiles can be constructed, including the i) left-hand translation distance profile, ii) left-hand rotation distance profile, iii) right-hand translation distance profile, and iv) right-hand rotation distance profile. These newly constructed profiles are normalized before being used to identify critical points.

Variance Characteristics-Based Profiles:

Relying solely on translation and rotation distances may lead to over-segmentation. To address this issue, the variance characteristics-based approach can be used to capture the essential features of kinematic data [20]. 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. High variance points suggest significant changes in terms of angle or angular speed values. Understanding these variance characteristics can help differentiate between continuous motion and transition periods, thus enhancing the segmentation accuracy.

Assume there are N new frames generated from random translation matrix $^{\mathbf{n}}\mathbf{T}(t)$ and rotation matrix $^{\mathbf{n}}\mathbf{R(t)}$ separately for $n = 1, 2, \ldots, N$. The trajectory points in the new frames could be written as $^{\mathbf{n}}\mathbf{p}(t) = {}^{\mathbf{n}}\mathbf{R(t)}\mathbf{p}(t) + {}^{\mathbf{n}}\mathbf{T}(t)$. Given the rotation matrix $^{\mathbf{n}}\mathbf{R(t)}$, the Euler angles $\mathbf{e}(t)$ in the new rotation frame for each n can be derived. The variance characteristics-based profile can be calculated based on all new frames at each time step, as illustrated in (1):

$$\begin{aligned} Var^n_{trans}(t) &= Var[^n p_x(t)] + Var[^n p_y(t)] + Var[^n p_z(t)] \\ Var^n_{rot}(t) &= Var[^n e_x(t)] + Var[^n e_y(t)] + Var[^n e_z(t)] \end{aligned} \tag{1}$$

Four variance characteristics-based profiles can be constructed, including the i) left-hand translation variance profile, ii) left-hand rotation variance profile, iii) right-hand translation variance profile, iv) right-hand rotation variance profile.

*2) **Critical Points** Determination:* Since segmentation points separate two gestures, they typically occur when there is a significant distance between two consecutive points along a trajectory. That is to say, the critical points normally appear at the corners of the peaks among the newly constructed characteristics profiles. They can thus be identified when two points have a large distance among all the profiles. All the critical points obtained from different profiles can be combined as a point set, which can be later used for potential segmentation points identification.

Two methods can be used to identify these peaks and corners for determining the critical points. The first method involves specifying a threshold height and prominence, while the second method involves using continuous wavelet transformation (CWT) to identify relative peaks and corners. Both methods have been widely used in the literature for identifying critical points in various types of data. The comparisons of these two methods can be found in the supplementary materials on our website.

C. *Pre-Segmentation Based on Vision Data*

1) Transformer-Based Feature Extraction: Deep neural networks have demonstrated their immense capabilities in the field of computer vision. More recently, Transformers have been proven to outperform CNNs and RNNs when they are applied to sequential data processing, since they can deal with long-range context dependencies [21]. Transformers have the capacity to establish temporal links between the present and past frames based on self-attention mechanism [22]. Therefore, a modified Transformer-based architecture (see Fig. 2) is integrated with one of the Deep Residual Neural Network model-based backbones to implement visual feature extraction in this paper. Specifically, the ResNet-18 architecture with pre-trained weights from ImageNet is used as the backbone for the model, which accelerates the model training process. The remaining weights in the architecture are fine-tuned using the limited labeled data in JIGSAWS [21].

To prepare for model training, the video data is initially transformed into a sequence of images. For training the Transformer-based model, every input consists of a series of l images that share the same gesture label. The value of l can be referred to as the sliding window size, representing the length of a sequence of images used as model input. All inputs have $(l - s)$ overlapping frames, where s represents the step size for sequential data generation. In this paper, $l = 30$ and $s = 5$ are used to construct a new dataset with a reorganized data structure for model fine-tuning. After extracting the features from the vision data, we obtain a new dataset that includes sequences of labeled data represented by feature vectors.

2) Feature Clustering Based on Unsupervised Learning: The feature clustering method used in this work is inspired by [12], where an unsupervised Transition State Clustering (TSC) algorithm was used to identify potential segmentation points between two Gaussian clusters. It has been demonstrated that DP-GMM, learning through Expectation-Maximization, exhibits excellent performance in clustering

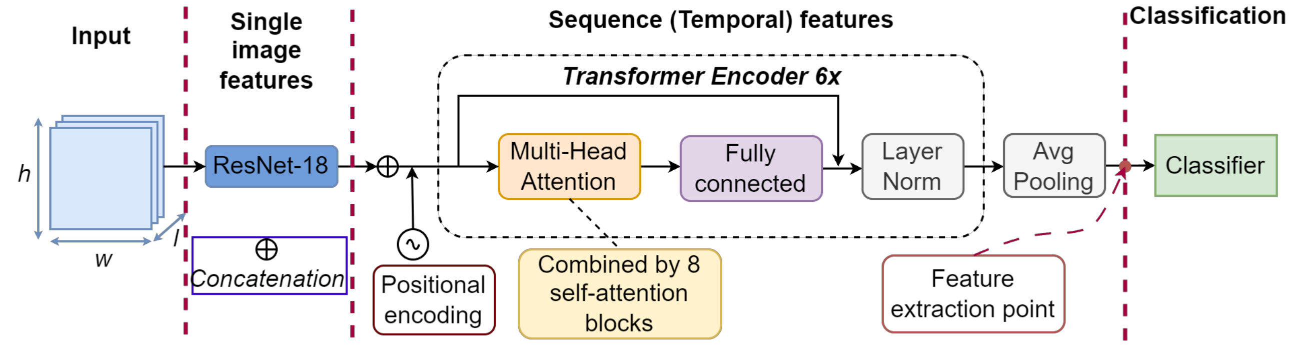

Fig. 2. Network Architecture of the Transformer-based network with a pre-trained 'ResNet-18' backbone for visual features extraction.

high-dimensional data without prior knowledge of the ground truth cluster number.

If there is a significant difference between two adjacent features, it is likely that the gestures change at that point, and thus the potential segmentation points could be found. The visual critical points extraction is performed using two layers of DP-GMM. The first layer clusters across all the frames to find as many critical points as possible, while the second layer further clusters those critical points to determine the potential segmentation points.

For the DP-GMM method, Dirichlet Process is a probability distribution that resolves the issue of requiring a predetermined number of classes when implementing a pure GMM. Suppose there are N data points $[x_1, x_2, ..., x_N]$, where each is generated from a different distribution $g_i = g_1, g_2, ..., g_N$, and each distribution has corresponding parameters $\theta_i = \theta_1, \theta_2, ..., \theta_N$. Assuming each g_i follows a distinct Gaussian distribution, θ_i follows a certain continuous distribution $H(\theta)$. The Dirichlet Process involves constructing a discrete distribution G to make $\theta_i \sim G$, where $G \sim DP(\alpha, H)$. Here, α is known as the concentration parameter, which controls the shape of the distribution.

D. *Final Segmentation Point Identification*

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm that can be used to cluster points with sufficient density. Unlike other clustering algorithms, it does not require the number of clusters to be specified in advance and only requires two hyperparameters. DBSCAN is also capable of identifying and handling abnormal values, which is effective in eliminating misidentified critical points or potential segmentation points.

In this paper, two layers of DBSCAN were utilized to locate segmentation points. The first layer is used to cluster all the kinematic critical points and therefore obtain kinematic potential segmentation points. The second layer combines the output of both kinematic and vision potential segmentation points to identify the final segmentation points. To implement this algorithm, two parameters must be defined: the minimum number of points ($minp$) required to determine a dense region, and the maximum distance (eps) between two points. Potential and final segmentation points should have a sufficient number of other data points (defined by $minp$) within a specified distance (defined by eps). Boundary points refer to data points that are not core points but lie within the neighborhood of a core point. If a point does not belong to either of these categories, it is classified as a noise point.

III. Experiments and Results Analysis

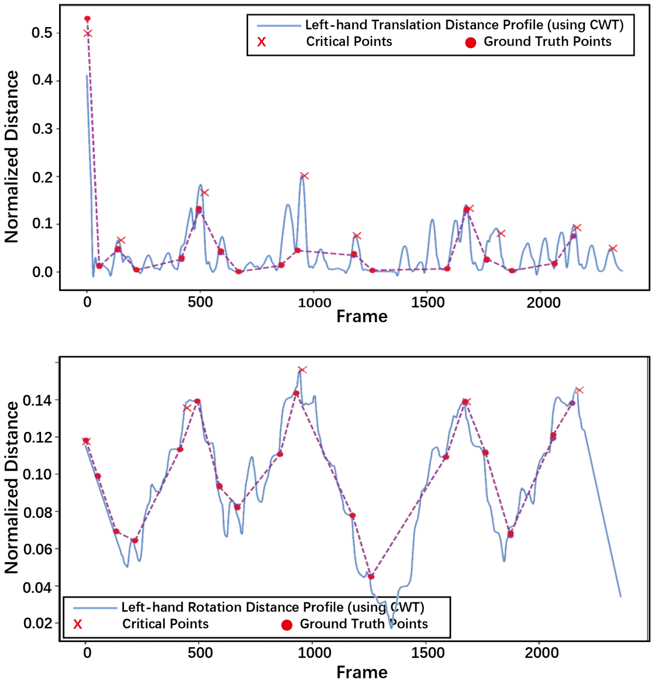

Fig. 3. Critical points determined based on the left-hand translation distance profile and left-hand rotation distance profile. (Data Source: Suturing File F004 in JIGSAWS)

A. *Experiment Design*

The dataset used in this study consists of a total of 103 demonstration files, which are divided into a training set, a validation set, and a test set in a 69:21:13 ratio. During the model training process for the Transformer network, images are cropped to 80x80 pixels to accelerate the network training process and accommodate CUDA memory. To prevent overfitting, we adopted data augmentation techniques such as random rescaling, random cropping, and random rotation within a range of -15 to 15 degrees [21]. The batch size is set to 30, and the total number of training epochs is 80. We use the Adam optimizer with weight decay, and the loss function is constructed using categorical cross-entropy. The initial learning rate is set to 0.001

Bayesian optimization is used to automatically find the optimized values of hyperparameters [23]. Specifically, the concentration parameters for the two DP-GMM layers (denoted as $\alpha 1$ and $\alpha 2$), the component numbers (n_1 and n_2) in the two Dirichlet Processes, and the maximum distance *eps* for DBSCAN can be fine-tuned through Bayesian optimization. The F1-score is chosen as the target function for optimization. The results of the optimized hyperparameters found through Bayesian Optimization are summarized in Table I.

TABLE I: Optimized Hyperparameters Determined Based on Bayesian Optimization

Task	n_1	n_2	$\alpha 1$	$\alpha 2$	*eps*
Suturing	768	145	9.437	3.729	28
Needle Passing	545	708	183.5	4.489	15.49
Knot Tying	658	445	157.8	162.2	31.76

Three standard frame-wise metrics (precision, recall, and F1-score) are used to evaluate the segmentation results [7]. As for multi-class classification, the micro and the macro average are calculated based on the confusion matrix as shown in the following equations.

$$Micro_{precision} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{2}$$

$$Micro_{recall} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{3}$$

$$Macro_{precision} = \frac{1}{n} \sum_{i=1}^{n} P_i \tag{4}$$

$$Macro_{recall} = \frac{1}{n} \sum_{i=1}^{n} R_i \tag{5}$$

where TP_i, FP_i,FN_i denotes True Positive, False Positive, and False Negative for each class i, P_i, R_i denotes the precision and recall for each class i.

B. Kinematics-Based Segmentation

Smooth trajectory profiles can be obtained after applying the Kalman filter, as oscillation points are eliminated, and critical points can be identified with higher accuracy. We conducted follow-up experiments and discovered that critical points were more accurately identified after applying the Savitzky-Golay filter.

Fig. 3 provides an example of critical points extracted from the left-hand translation and rotation distance profiles using CWT. A comparison of the critical segmentation points obtained from the two profiles suggests that the rotation distance profile provides more accurate results than the translation distance profile. This may be due to the fact that the Suturing Task involves more rotational motion than translational motion, resulting in more prominent critical points in the rotation distance profile. Critical points identified from the rotation variance profile are shown in Fig. 4 as examples.

By combining all the **critical points** obtained from both distance characteristics-based and variance characteristics-based profiles, we can further obtain a set of **potential segmentation points** after applying the first layer of DBSCAN. Using these potential segmentation points, we assessed the kinematics-based segmentation with the following metrics: i) mean recall, ii) precision, and iii) F1 score. For the Knot Tying task, these metrics are 0.482, 0.251, and 0.330, respectively. Similarly, for the Suturing Task, the mean recall, precision, and F1 score are 0.488, 0.242, and 0.324, respectively. For the Needle Passing task, the corresponding values are 0.372, 0.145, and 0.210, respectively.

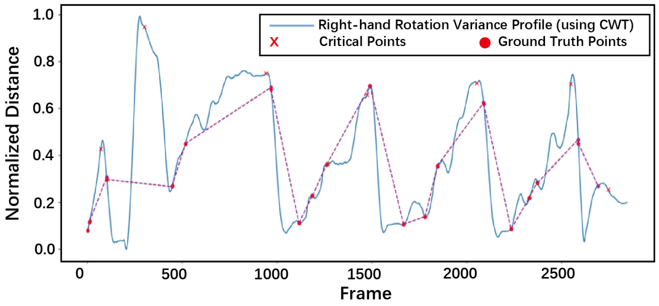

Fig. 4. An example of right-hand rotation variance profile (Data Source: Suturing File B005, JIGSAWS)

The aforementioned results reveal that though the proposed method can correctly identify many segmentation points, there is still a noticeable deviation between the ground truth segmentation points and the potential segmentation points when relying solely on kinematics data. One possible explanation is that the dataset is captured by surgeons with varying robotic surgical experience. Some surgeons' operation data may has a significant smoothness discrepancy in the trajectory. The smoother the trajectory is, the shorter pause it might contain between different gestures, and there might not be noteworthy critical points. Thus, the use of visual data is necessary to help improve segmentation accuracy, which contains some important context information to help differentiate different gestures.

C. Vision-Based Segmentation Results

The testing accuracy of the Transformer-based model on the test dataset is 0.93. The mean of recall, precision, and F1-score for all Suturing segmentation metrics are 0.516, 0.634, and 0.550, respectively. As for Needle Passing and Knot Tying task, the scores for all the evaluation metrics are summarized in Table II. Take the Knot Tying task as an example, the segmentation results based on vision data are shown in Fig. 5.

The accuracy of the segmentation points for the Needle Passing task was found to be the lowest among the three tasks. This could be attributed to the fact that most gestures for the Needle Passing task were conducted based on local operation, resulting in less significant differences in the distance characteristics-based profile compared to the other tasks. Additionally, critical features such as the needle pose and thread in the video data may be obscured, making the accurate identification of segmentation points challenging.

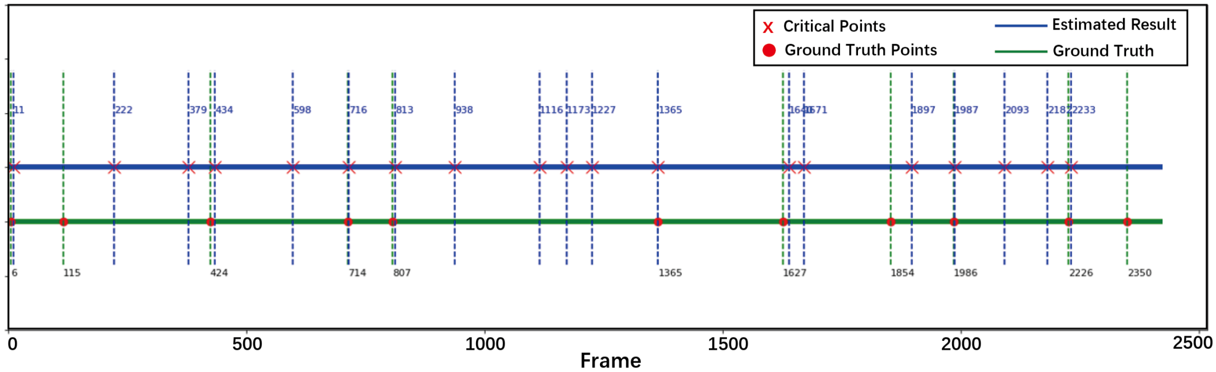

Fig. 5. An example of the segmentation result using vision data (Data Source: Knot Tying File I002, JIGSAWS)

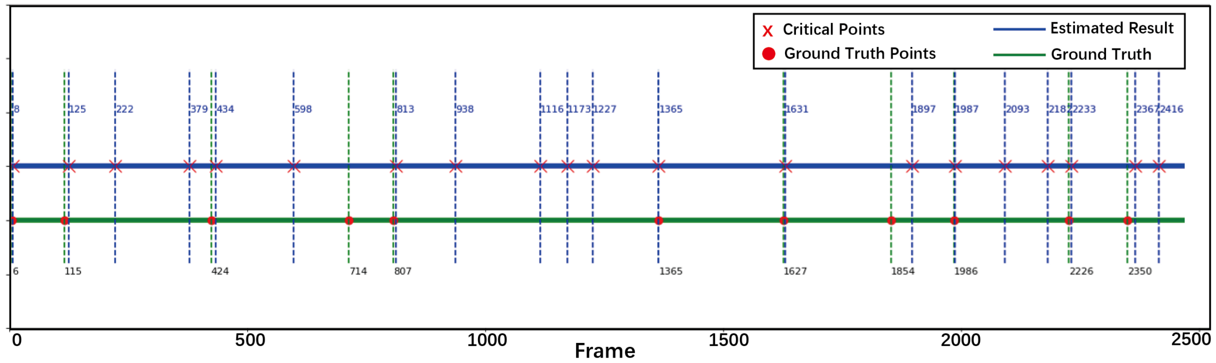

Fig. 6. An example of the segmentation result using vision and kinematics data. (Data Source: Knot Tying File I002, JIGSAWS)

TABLE II: Summary of Segmentation Results

Modality	Task	Recall	Precision	F1 score
Kinematics	Suturing	0.488	0.242	0.324
	Needle Passing	0.372	0.145	0.210
	Knot Tying	0.482	0.251	0.330
Vision	Suturing	0.516	0.634	0.559
	Needle Passing	0.415	0.620	0.488
	Knot Tying	0.567	0.673	0.615
kinematics + Vision	Suturing	0.533	0.770	0.630
	Needle Passing	0.510	0.647	0.570
	Knot Tying	0.569	0.780	0.657
	mean	0.537	0.745	0.623

D. Vision and Kinematics Fusion-Based Results

1) Final Segmentation Results: For the Needle Passing task, the average segmentation results in terms of recall, precision, and F1-score are 0.510, 0.647, and 0.570, respectively. For the Knot Tying task, the corresponding results are 0.569, 0.780, and 0.657, respectively. A comprehensive summary of all segmentation scores is presented in Table II. These findings illustrate the superior performance of our proposed method in surgical gesture segmentation, as our framework achieves improvements in both recall and precision.

2) Final Classification Results: The segmentation results are utilized to partition the entire video into segments, generating a new dataset that enables the classification of each segment for the surgical gesture recognition task. The results of this task are summarized in Table III. Fig. 7 provides examples of classification results with estimated segmentation points. This indicates that the Transformer-based model can be used to refine the results generated by the unsupervised methods and improve the classification accuracy.

Based on the analysis of the confusion matrix (see supplementary materials provided by our website), we observed that Gesture 10 (Loosening more suture) can be easily identified. Conversely, Gesture 5 (Moving to center with needle in grip) has the lowest classification accuracy, with 25% of Gesture 5 instances being misidentified as either Gesture 1 (Reaching for the needle with the right hand) or Gesture 2 (Positioning the needle). A potential reason for this could be that the Transformer-based model can only accept 30 frames as synchronous input, and different gestures might involve similar motion within these 30 frames. Since Gestures 1 and 2 involve the state of 'having the needle in grip' at the center of the image, they might be misclassified as Gesture 5.

TABLE III: Summary of Classification Results

Metrics	Result	Metrics	Result
Recall_micro	0.868	Precision_micro	0.868
Recall_macro	0.829	Precision_macro	0.830
Recall_weighted	0.867	Precision_weighted	0.879
Accuracy	0.856		

E. Comparisons with State-of-the-Art Methods

LSTM [24] and TCN [25] are considered state-of-the-art supervised learning methods for classification. In addition, Bimanual space and variance [20], Soft-UGS [26], and Zero-shot [27] represent state-of-the-art unsupervised learning techniques.

In this paper, these methods serve as baselines for comparative studies and are all evaluated using the Suturing Task. The results presented in Table IV, reveal that our proposed

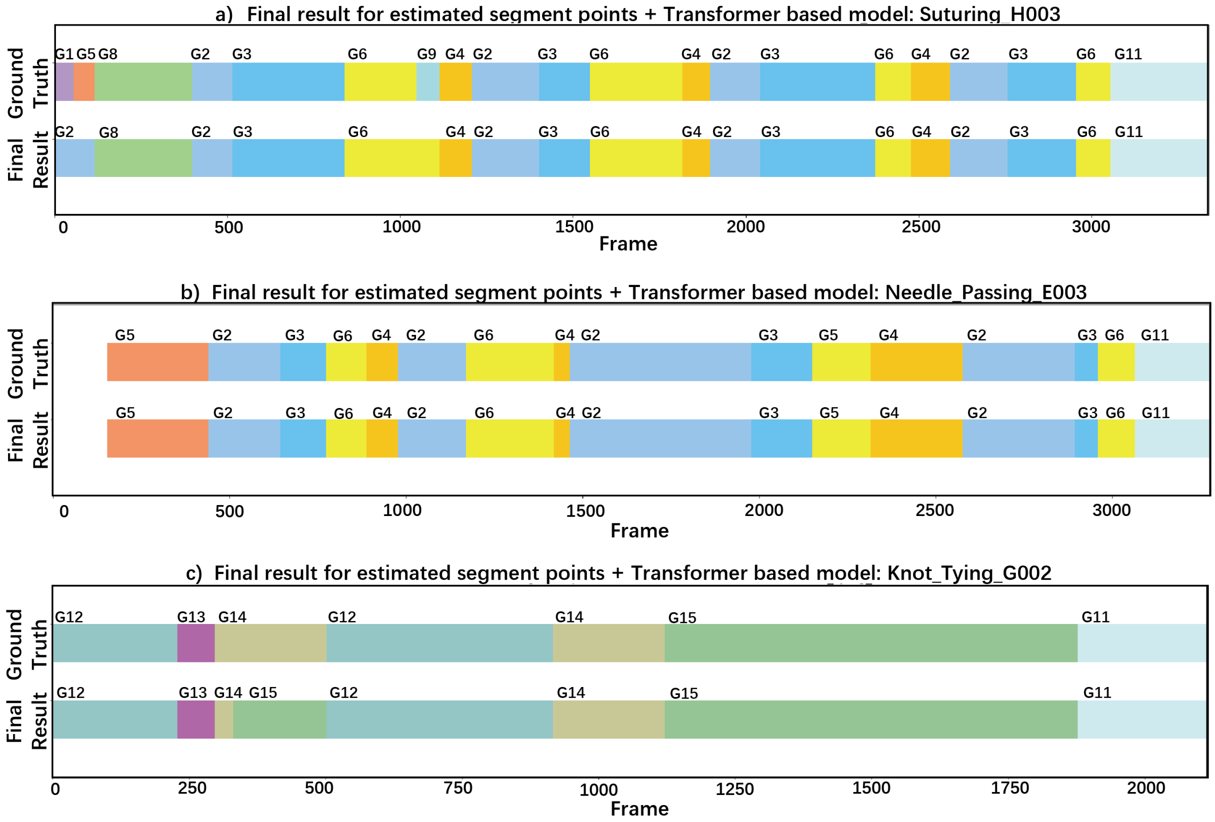

Fig. 7. Examples of surgical gesture classification results.

TABLE IV: Comparisons with other State-of-the-Art Approaches

Method	Metrics	Scores
Space and variance	Recall/Precision/F1	0.652/ 0.92/0.754
Soft-UGS	Recall/Precision/F1	0.74/0.71/0.72
LSTM	Acc	0.805
TCN	Acc	0.796
Zero-shot	Acc	0.56
CNN+LC-Sc-CRF	Acc	0.766
Ours	Acc	**0.856**

method outperforms the baseline techniques in terms of accuracy.

IV. Discussions and Future Work

We propose a hierarchical semi-supervised learning framework for surgical gesture segmentation and recognition. 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. By leveraging implicit information from both kinematic and video data, the proposed method is capable of successfully identifying segmentation points and recognizing gesture labels. The effectiveness of our proposed method was verified using the JIGSAWS dataset, achieving an impressive accuracy of 0.856 for classification and an F1 score of 0.623 for segmentation.

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 [28]. Furthermore, we aim to apply the proposed method to support the development of automation in robotic surgery based on learning from demonstration.

References

[1] D. Zhang, W. Si, W. Fan, Y. Guan, and C. Yang, “From teleoperation to autonomous robot-assisted microsurgery: A survey,” *Machine Intelligence Research*, vol. 19, no. 4, pp. 288–306, 2022.

[2] R. Wang, D. Zhang, Q. Li, X.-Y. Zhou, and B. Lo, “Real-time surgical environment enhancement for robot-assisted minimally invasive surgery based on super-resolution,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 3434–3440.

[3] L. Maier-Hein, M. Eisenmann, D. Sarikaya, K. März, T. Collins, A. Malpani, J. Fallert, H. Feussner, S. Giannarou, P. Mascagni *et al.*, “Surgical data science–from concepts toward clinical translation,” *Medical image analysis*, vol. 76, p. 102306, 2022.

[4] D. Zhang, B. Xiao, B. Huang, L. Zhang, J. Liu, and G.-Z. Yang, “A self-adaptive motion scaling framework for surgical robot remote control,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 359–366, 2018.

[5] D. Zhang, Z. Wu, J. Chen, A. Gao, X. Chen, P. Li, Z. Wang, G. Yang, B. Lo, and G.-Z. Yang, “Automatic microsurgical skill assessment based on cross-domain transfer learning,” *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4148–4155, 2020.

[6] J. Chen, D. Zhang, A. Munawar, R. Zhu, B. Lo, G. S. Fischer, and G.-Z. Yang, “Supervised semi-autonomous control for surgical robot based on bayesian optimization,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 2943–2949.

[7] B. van Amsterdam, M. J. Clarkson, and D. Stoyanov, "Gesture recognition in robotic surgery: a review," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 6, 2021.

[8] M. K. Bhuyan, D. Ajay Kumar, K. F. MacDorman, and Y. Iwahori, "A novel set of features for continuous hand gesture recognition," *Journal on Multimodal User Interfaces*, vol. 8, no. 4, pp. 333–343, 2014.

[9] L. Tao, L. Zappella, G. D. Hager, and R. Vidal, "Surgical gesture segmentation and recognition," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2013, pp. 339–346.

[10] D. Zhang, R. Wang, and B. Lo, "Surgical gesture recognition based on bidirectional multi-layer independently rnn with explainable spatial feature extraction," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1350–1356.

[11] R. DiPietro, N. Ahmidi, A. Malpani, M. Waldram, G. I. Lee, M. R. Lee, S. S. Vedula, and G. D. Hager, "Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks," *International journal of computer assisted radiology and surgery*, vol. 14, no. 11, pp. 2005–2020, 2019.

[12] S. Krishnan, A. Garg, S. Patil, C. Lea, G. Hager, P. Abbeel, and K. Goldberg, "Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1595–1618, 2017.

[13] A. Murali, A. Garg, S. Krishnan, F. T. Pokorny, P. Abbeel, T. Darrell, and K. Goldberg, "Tsc-dl: Unsupervised trajectory segmentation of multi-modal surgical demonstrations with deep learning," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 4150–4157.

[14] H. Zhao, J. Xie, Z. Shao, Y. Qu, Y. Guan, and J. Tan, "A fast unsupervised approach for multi-modality surgical trajectory segmentation," *IEEE Access*, vol. 6, pp. 56 411–56 422, 2018.

[15] Y. Gao, S. S. Vedula, G. I. Lee, M. R. Lee, S. Khudanpur, and G. D. Hager, "Unsupervised surgical data alignment with application to automatic activity annotation," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 4158–4163.

[16] R. DiPietro and G. D. Hager, "Automated surgical activity recognition with one labeled sequence," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2019, pp. 458–466.

[17] B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," *IET Computer Vision*, vol. 12, no. 1, pp. 3–15, 2018.

[18] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI workshop: M2cai*, vol. 3, 2014, p. 3.

[19] Y. Guo, Y. Li, and Z. Shao, "Rrv: A spatiotemporal descriptor for rigid body motion recognition," *IEEE transactions on cybernetics*, vol. 48, no. 5, pp. 1513–1525, 2017.

[20] Y.-Y. Tsai, Y. Guo, and G.-Z. Yang, "Unsupervised task segmentation approach for bimanual surgical tasks using spatiotemporal and variance properties," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1–7.

[21] A. D'Eusanio, A. Simoni, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara, "A transformer-based network for dynamic hand gesture recognition," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 623–632.

[22] T. Czempiel, M. Paschali, D. Ostler, S. T. Kim, B. Busam, and N. Navab, "Opera: Attention-regularized transformers for surgical phase recognition," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 604–614.

[23] P. I. Frazier, "A tutorial on bayesian optimization," *arXiv preprint arXiv:1807.02811*, 2018.

[24] R. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, and G. D. Hager, "Recognizing surgical activities with recurrent neural networks," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 551–558.

[25] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European conference on computer vision*. Springer, 2016, pp. 47–54.

[26] M. J. Fard, S. Ameri, R. B. Chinnam, and R. D. Ellis, "Soft boundary approach for unsupervised gesture segmentation in robotic-assisted surgery," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 171–178, 2016.

[27] T. S. Kim, J. Jones, M. Peven, Z. Xiao, J. Bai, Y. Zhang, W. Qiu, A. Yuille, and G. D. Hager, "Daszl: Dynamic action signatures for zero-shot learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 1817–1826.

[28] D. Zhang, Z. Wu, J. Chen, R. Zhu, A. Munawar, B. Xiao, Y. Guan, H. Su, W. Hong, Y. Guo *et al.*, "Human-robot shared control for surgical robot based on context-aware sim-to-real adaptation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7694–7700.