Surgical Gesture Recognition Based on Bidirectional Multi-Layer Independently RNN with Explainable Spatial Feature Extraction

Dandan Zhang[1*], *Student Member, IEEE*, Ruoxi Wang[1*], Benny Lo[1], *Senior Member, IEEE*

***Abstract*— Minimally invasive surgery mainly consists of a series of sub-tasks, which can be decomposed into basic gestures or contexts. As a prerequisite of autonomic operation, surgical gesture recognition can assist motion planning and decision-making, and build up context-aware knowledge to improve the surgical robot control quality. In this work, we aim to develop an effective surgical gesture recognition approach with an explainable feature extraction process.**

A Bidirectional Multi-Layer independently RNN (BML-indRNN) model is proposed in this paper, while spatial feature extraction is implemented via fine-tuning of a Deep Convolutional Neural Network (DCNN) model constructed based on the VGG architecture. To eliminate the black-box effects of DCNN, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed. It can provide explainable results by showing the regions of the surgical images that have a strong relationship with the surgical gesture classification results.

The proposed method was evaluated based on the suturing task with data obtained from the public available JIGSAWS database. Comparative studies were conducted to verify the proposed framework. Results indicated that the testing accuracy for the suturing task based on our proposed method is 87.13%, which outperforms most of the state-of-the-art algorithms.

I. Introduction

In the past few decades, Robot-Assisted Minimally Invasive Surgery (RAMIS) has transformed surgery and brought great benefits to patients, such as reduced recovery time and traumas [1]. Moreover, it enables surgeons to conduct surgery remotely via a master manipulator [2]–[4], which can protect surgeons from radiation and provide ergonomic comfort [5], [6]. With RAMIS, the surgical operation data can be recorded from the robotic system and used for surgical gesture segmentation and recognition, surgical skills assessment [7]–[9], robotic assistance [10]–[12], and automation [13]–[15]. Since surgical gesture recognition is essentially required for the support of high-level perception of surgical workflow [16], we focus on the implementation of automatic surgical gesture recognition with explainable features in this paper.

To process sequential data for surgical gesture recognition, classical approaches, such as Hidden Markov Model (HMM) [17] has been widely used. HMM estimates the transition of hidden states to characterize the surgical task sequences. Surgical gestures can be modeled as one or more states of an HMM, while the observations can be modeled differently [18]. For example, the observations are assumed to be generated from a lower-dimensional latent space using Factor Analyzed HMMs (FA-HMMs) and Switched Linear Dynamical Systems (SLDSs) [19]. Gaussian mixture model (GMM) [16], [20] and mixtures of factor analyzers (MFAs) can be used to capture the variability of complex operations as observations. In addition to HMM, Dynamic Time Wrapping [21], Conditional Random Fields (CRF) [22] and other machine learning based time sequences processing technologies have been utilized for surgical pattern recognition as well.

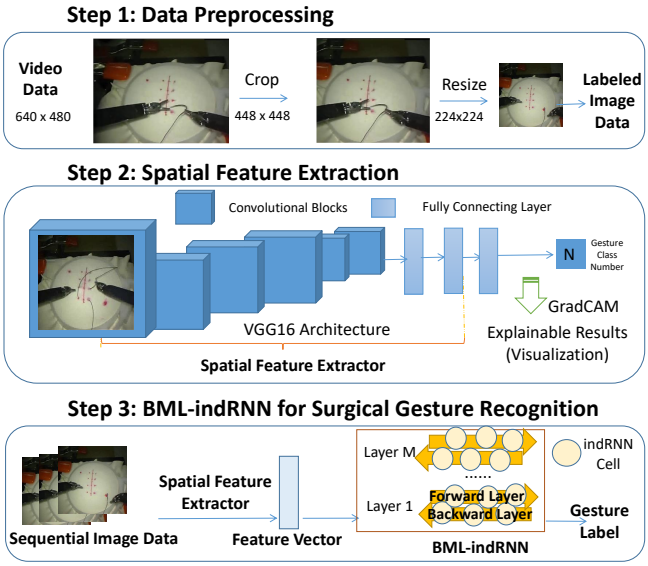

Fig. 1. The workflow for the implementation of the proposed method for surgical gesture classification.

Deep Neural Networks (DNNs) have emerged with promising results in many applications. Compare to most of the traditional statistical models for sequential data modeling, DNNs can extract significant features from data automatically without manually selecting features [23], [24]. To incorporate temporal information for analyzing time series data, Long Short-Term Memory (LSTM) has been employed, which can preserve temporal characteristics of the signals [25] by fusing information from previous steps and instantaneous inputs. In addition, Temporal convolutional network (TCN) [26] and 3D convolutional network [27] have been proved to be promising for sequential data processing as well, which have been used to extract representative visual features for surgical gesture recognition. They have shown an important role in applications such as action recognition, semantic labeling, language or video processing. More recently, Fusion-KV [28] was proposed to combine the vision

* These authors contribute to this paper equally. [1]The Hamlyn Centre for Robotic Surgery, Imperial College London, London SW7 2AZ, United Kingdom. Corresponding Author email: d.zhang17@imperial.ac.uk

data and kinematics data recorded from the surgical robotic system for gesture recognition, while Fusion-KV with an attention-based LSTM decoder has been utilized to further enhance the recognition performance [29].

In addition to supervised learning based approaches, deep reinforcement learning (RL) has been investigated in gesture recognition, where an agent learned its policy of surgical gesture segmentation and classification by interacting with the surgical data [30]. However, the network confusion problem has not been addressed for some less frequent surgical gestures [31]. A reinforcement learning and tree search based framework has been proved to be effective for surgical gesture recognition [32], where the visual features are regarded as the environment states while the determination of the classes for the surgical gesture is regarded as actions. A tree search algorithm is used to combine the outputs of the policy and the value network to obtain a better performance.

The aim of this paper is to propose a Bidirectional Multi-Layer independently RNN (BML-indRNN) model with spatial feature extraction for surgical gesture recognition. Apart from accurate surgical gesture recognition, the work also provides explainable results by showing how the neural network classifies the surgical gestures. Considering that video data includes semantic information and can be more intuitive for identifying the context compared to kinematics data, we use video data for the implementation of automatic surgical gesture recognition.

The details of the paper are described as follows. Firstly, the methodology is introduced in Section II. Secondly, the experimental setup and results analysis are presented in Section III. Finally, conclusions are drawn in Section IV.

II. Methodology

A. Overview

Video data of surgical operation includes both spatial and temporal features that are useful for surgical gesture recognition. Deep Convolutional Neural Network (DCNN) has demonstrated promising results in image classification with powerful spatial feature extraction functions. However, it is not suitable for identifying the temporal features of sequential data. Therefore, we intend to combine the advantages of DCNN for spatial feature extraction and RNN for temporal feature extraction, through which the method can enhance the gesture classification accuracy.

The overview of the workflow of the proposed method for surgical gesture classification is shown in Fig. 1. The first step is data pre-processing, which requires transforming video data into a series of sequential images, cropping and reshaping the frames to the desired sizes and labeling each image to the corresponding surgical gesture class. The second step is spatial feature extraction, which is implemented by using the DCNN model. VGG model, as a typical architecture of DCNN is used for spatial feature extraction. It can compress the 2D image array to a 1D feature vector while preserving the most useful information before feeding into the RNN. The third step is surgical gesture recognition based on the proposed BML-indRNN model, which extracts the temporal features automatically and predicts the surgical gestures. More details of the processing pipeline are described in the following sections.

B. Database Description

Since the JIGSAWS database is a publicly available database with several surgical tasks, we chose the suturing task as an example to test the performance of the proposed method in this paper. Video data of the JIGSAWS database is used for validation. The definitions of the surgical gesture for suturing task are shown in Table. I, which are adapted from [33]. Ten types of surgical gestures are defined in total. Thus, surgical gesture recognition can be formulated as a 10-class classification problem.

TABLE I
Definitions for Surgical Gestures Involved in the Suturing Task

Surgical Task	Surgical Gesture Definition
Suturing	G1: Reaching for needle with right hand
	G2: Positioning needle
	G3: Pushing needle through tissue
	G4: Transferring needle from left to right
	G5: Moving to center with needle in grip
	G6: Pulling suture with left hand
	G7: Orienting needle
	G8: Using right hand to help tighten suture
	G9: Loosening more suture
	G10: Dropping suture at end and moving to end points

C. Explainable Spatial Feature Extraction

1) Model Construction: The JIGSAWS has been manually annotated with the ground-truth surgical gesture at frame level [34]. Therefore, the corresponding relationship between each frame and the gesture label can be obtained for model training. The size of the original frame of the video data is 640×480. The videos are recorded at a frame rate of $30hz$. During the model training process, the frame for the video data is downsampled with factor $r = 5$. Furthermore, each frame is cropped to have the dimension of 448×448 pixels and is resized to 224×224 pixels to accelerate the network training and inferencing process.

VGG16 model is designed for image classification [35], which has been proven to be able to classify images accurately from the ImageNet dataset [36]. To extract meaningful features for the state representation of the surgical operation image data, we use the VGG16 model with pre-trained weights on ImageNet for feature extraction. However, the pre-trained model may have a bias when used for feature extraction for surgical gesture classification, since the model is trained for object recognition originally. Therefore, we need to fine-tune the model for effective feature extraction using the task-orientated database to obtain a better state representation of the image data.

The modification of the architecture for spatial feature extraction is shown in Fig. 2. The weights of the convolutional blocks and the first fully connecting layer are frozen, which

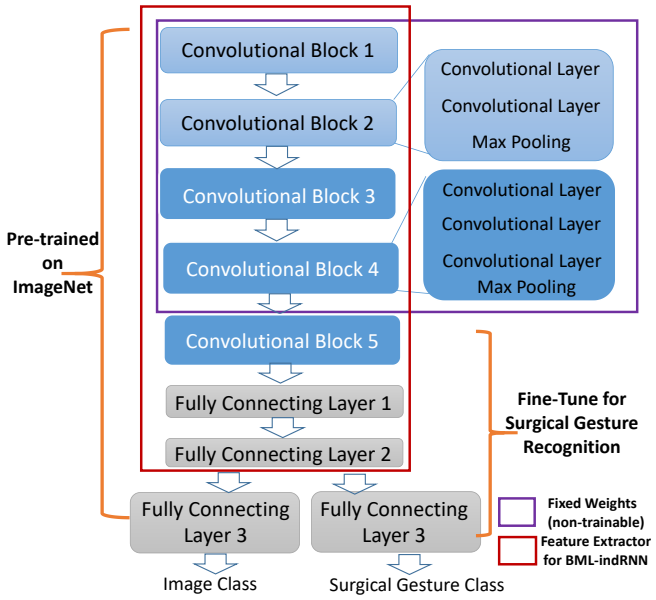

Fig. 2. The architecture of the feature extraction model.

can form a feature extraction model, while the final fully connecting layer of the network for mapping the features to specific classes of the ImageNet is removed. Following that, we add a fully connecting layer with input units of 4096×1 and output units of N to the end of the model, where N represents the number of surgical gesture classes defined for the specific recognition task.

During model training, the weights of the convolutional block 1 to convolutional block 4 are fixed, while the convolutional block 5 and all the fully connecting layers are trainable for frame-wise surgical gesture classification. Once the new model is obtained, a new feature extraction model can be constructed after removing the final fully connecting layer. The input of the feature extraction model is an image array after pre-processing, while the output of the model is a compact feature vector with the dimension of 4096×1. The obtained feature vector can then be fed into the BML-indRNN for surgical gesture classification, as described in the next section.

2) Interpretability: Though DCNN has been proven to have outstanding performance in classification, it has inherent black-box effects. In order to make the model transparent for people to under the reason why the series of frames are classified as a specific gesture, visualization techniques should be used to provide interpretable results obtained via the model. In this way, the effectiveness of the spatial feature extraction process can be evaluated qualitatively.

Class Activation Map has been developed to show the critical parts that contribute the most to the classification results [37] [38]. However, a global activation map layer in the neural network model is necessary, which limits its general applications. Gradient-weighted Class Activation Mapping (Grad-CAM) [39] has been proposed, which is able to mitigate the black-box effect of DCNN without re-training the model. The class-specific gradient information can be conveyed to the final convolutional layer of a DCNN, while the critical parts that contribute the most to the classification results can be visualized in the form of a coarse localization map with important regions highlighted. Grad-CAM can help users establish trust in predictions from DCNN. Therefore, Grad-CAM is used to provide an explainable spatial feature extraction post-hoc analysis in this paper.

D. BML-indRNN

In most of the RNN based methods for sequential data processing, LSTM units are widely used. For the applications of surgical workflow analysis, the length of the surgical operation videos normally has more than 1000 frames. Under this condition, traditional RNN methods with LSTM units may suffer from gradient decay over layers. Therefore, an improved RNN based model is required to develop for surgical gesture recognition.

The concept of an Independently recurrent neural network (IndRNN) has been proposed in [40]. The hidden state $\mathbf{h_{t-1}}$ and $\mathbf{h_t}$ at timestep $t-1$ and t are independent of each other, which provides better interpretability of the spatial features ($\mathbf{W}$) and temporal features ($\mathbf{u}$). Multiple layers can be stacked together to conduct correlation among different neurons. Therefore, each layer can process the outputs of all the neurons in the previous layer.

We aim to apply the Independently RNN (IndRNN) cell [40] to construct the model of BML-indRNN. The IndRNN cell replaces the matrix multiplication (Eq. 1) by element-wise vector multiplication to calculate the hidden state, which means that each neuron has a single recurrent weight connected to its last hidden state.

$$\mathbf{h_t} = \sigma(\mathbf{W x_t} + \mathbf{U h_{t-1}} + \mathbf{b}) \tag{1}$$

The mathematical expression of the IndRNN is as follows:

$$\mathbf{h_t} = \sigma(\mathbf{W x_t} + \mathbf{u} \odot \mathbf{h_{t-1}} + \mathbf{b}) \tag{2}$$

where $\odot$ represents Hadamard product, while the traditional operation is a dot product. $\mathbf{x_t} \in \mathbb{R}^M$, $\mathbf{h_t} \in \mathbb{R}^N$ are the input and hidden state at time step t. $\mathbf{W} \in \mathbb{R}^{N \times M}$, $\mathbf{u} \in \mathbb{R}^N$, and $\mathbf{b} \in \mathbb{R}^N$ are the weights for the current input and the recurrent input, and the bias of the neurons. $\sigma(.)$ is an element-wise activation fuction of the neurons, and $\mathbf{N}$ is the number of neurons in the RNN layer.

To further improve the performance of the neural network for surgical gesture recognition, we extend the original independent RNN concept to a Bidirectional Multi-Layer indRNN (BML-indRNN) model. The temporal features are considered when the sequence is reversed. The output at time t depends not only on the information at the previous moment but also on the future moment. The basic concept is to put two IndRNNs together by feeding the input time-sequential data in normal time order for one network and in reverse time order for another, which utilizes both forward and backward information about the sequence at every time step. In this way, the neural network can extract features in the forward and reverse order respectively. The architecture

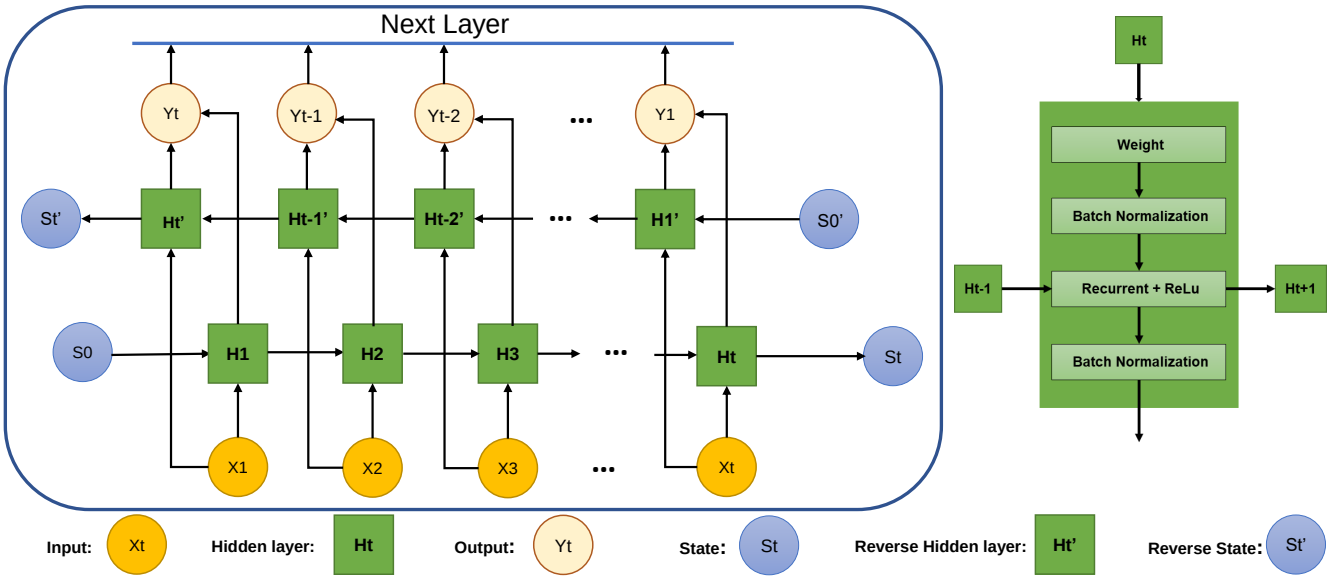

Fig. 3. The proposed BML-indRNN model for surgical gesture recognition.

of the BML-indRNN model is shown in Fig. 3. BML-indRNN model is constructed by three layers, including 64 hidden units in this paper.

III. Experiments Results and Analysis

A. Evaluation Metrics

Suppose that there are n targeted classes, $C = [i, j]$ represents the confusion matrix with the dimension of $n \times n$. Each element of the confusion matrix represents that the sample from class i is predicted as class j. Let TP denotes True Positive, TN denotes True Negative, FP denotes False Positive, FN denotes False Negative in the classification setting. Accuracy (acc), precision ($precision$), recall ($recall$) and F1-score (F_1) can be obtained as follows.

$$\begin{cases} acc = (TP + TN)/(TP + FN + FP + TN) \\ precision = TP/(TP + FP) \\ recall = TP/(TP + FN) \\ F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \end{cases} \tag{3}$$

A confusion matrix can be generated for the evaluation of the multi-class classification setting, while Micro average and Macro average can be computed as follows for evaluations of the proposed method.

$$\begin{cases} Micro = \frac{\sum_{i=1}^{n} C[i,i]}{\sum_{i,j=1}^{n} C[i,j]}; \\ Macro = \frac{1}{n} \sum_{i=1}^{n} \frac{C[i,i]}{\sum_{j=1}^{n} C[i,j]} \end{cases} \tag{4}$$

The Micro average is computed as the average of total correct predictions across all classes. Macro average represents the average TP rates for each class. Macro F1-score (F_{m1}) is defined as the mean of class-wise F1-scores, while Macro Recall and Macro precision can be obtained for evaluation as well.

B. Experimental Results

The proposed algorithm was implemented in Python using TensorFlow and Keras, and was trained on a PC with a GeForce GTX 1050 GPU (NVidia Corporation) and 8 GB of RAM.

1) Results for Spatial Feature Extraction Model Training: The spatial feature extraction model is trained at first, with SGD as the optimizer and the loss function constructed by categorical cross-entropy. The batch size is set as 10. The learning rate decreases during the model training process, an exponential decay function is applied to reduce the learning rate as the training progresses.

The data for experimental validation is divided using Leave One Trial Out (LOTO) for all the experiments, the setup of which is the same as described in [33]. The testing database includes 37086 data points. As for the training database, 70% of data is used for training, while the remaining 30% of data is used for validation. The training accuracy is 0.977, while the testing accuracy is 0.693 for training the spatial feature extraction model.

2) Results for Interpretability: Fig. 4 demonstrates the explainable results for evaluating the Spatial Feature Extraction Model based on Grad-CAM. Image data selected from six different types of gesture classes is used as examples. For a particular category, the red color highlight the discriminative image regions used by the DCNN to identify that specific class of surgical gesture.

Comparisons are made between the explainable visualization results of the model with and without fine-tuning. It can be clearly seen that without fine-tuning, the highlighted area for determining the gesture class is random, which means that the DCNN generates the results by reasoning some areas that do not have a close relationship with the operation scene. As for the explainable visualization results demonstrated in the third row, it can be clearly seen that the highlighted areas

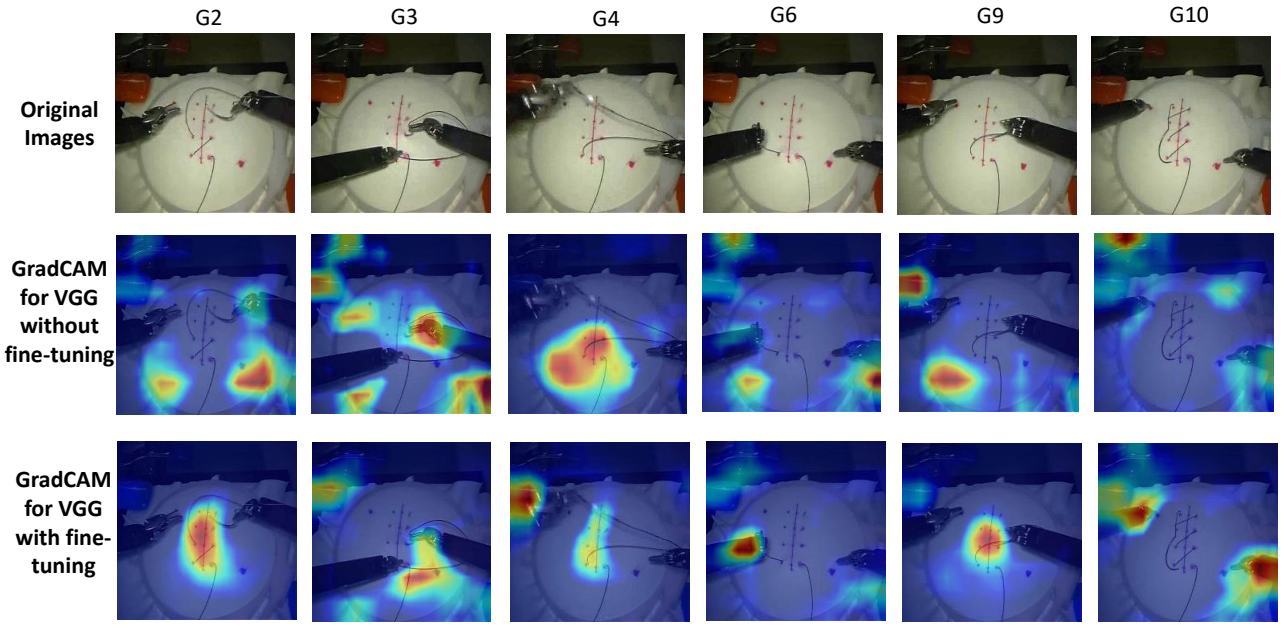

Fig. 4. Six examples of the analysis of the spatial feature extraction model using GradCAM and the comparison between model with and without fine-tuning.

are related to the states of the suture, the needles and the tooltip of the surgical tools, which have a close relationship with the contexts of operation scenes and the corresponding gesture type.

TABLE II
SURGICAL GESTURE RECOGNITION RESULTS BASED ON THE BML-INDRNN MODEL

	Without Tuning	With Tuning
Micro	83.15%	88.95%
Macro ± std	73.41%±19.70%	86.40%±8.38%
Precision ± std	69.68%±2.49%	84.59%±1.20%
Macro Recall	0.667	0.864
Macro F1-Score	0.666	0.849

3) Results for BML-indRNN Model Training: The results for surgical gesture recognition using the proposed BML-indRNN are summarized in Table II, while the confusion matrix for the suturing task is calculated, as shown in Fig. 5. The hyperparameters can be tuned for different surgical tasks to reach the desired performance.

To verify the effectiveness of the spatial feature extraction model, comparisons are made between with and without using the spatial feature extraction model as well as with and without fine-tuning. The color-coded ribbon illustration for the comparisons between the ground-truth data and the prediction results are shown in Fig. 6. Different colors represent different surgical gestures in the entire surgical procedure. The top line represents the ground-truth and the bottom line represents the predicted results.

The 2D image array obtained from video can be reshaped to a 1D vector and then be fed into the BML-indRNN model for training. In this case, no feature extraction procedure is

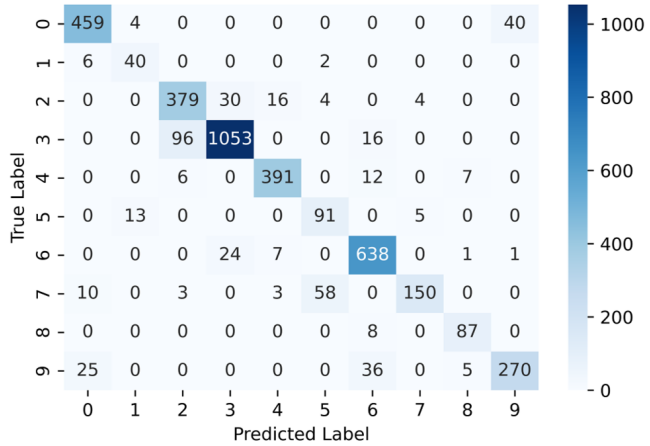

Fig. 5. The confusion matrix for the surgical gesture classification results of the suturing task.

conducted for model training, while the prediction results are shown in Fig. 6 (a), which can be regarded as the baseline for a comparative study. As indicated by the results in Fig 6 (b) and (c), we can conclude that fine-tuning can improve surgical gesture recognition performance.

4) Comparisons with the State-of-the-Art: Markov/semi-Markov conditional random field (MsM-CRF) model [41], Segmental spatiotemporal convolutional neural network (Seg-ST-CNN) [42], Temporal convolutional networks (TC-N) [26], and together with Deep Reinforcement Learning (DRL) [30] have been evaluated on the suturing task using JIGSAWS. All these methods can be used as the baselines for the performance evaluation of the surgical gesture recognition for the suturing task.

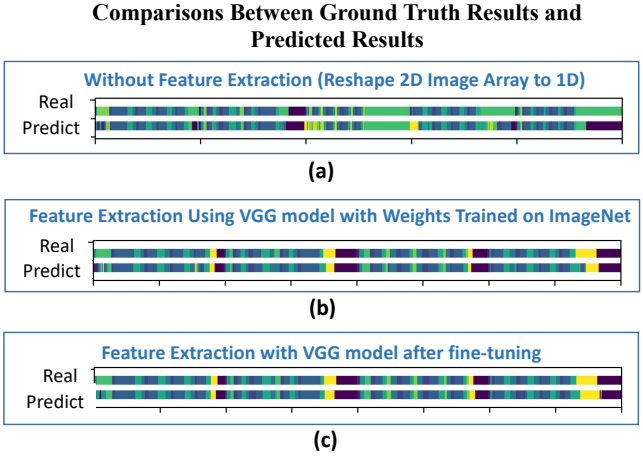

Fig. 6. The visualization of comparisons between ground-truth data and predicted results. (a) Results predicted by model trained without DCNN based feature extraction. (b) Results predicted by model trained with VGG based model without fine-tuning. (c) Results predicted by model trained with VGG based model with fine-tuning.

TABLE III

THE RESULTS FOR COMPARISONS BETWEEN THE PROPOSED METHODS AND THE BASELINE METHODS.

Method	Accuracy	Method	Accuracy
MsM-CRF	0.718	Seg-ST-CNN	0.742
TCN	0.814	TCN+Deep RL	0.814
BML-indRNN	0.822	**BML-indRNN+DCNN**	0.871

Table. III shows the results of the comparisons between the state-of-the-art algorithms and the proposed method in terms of the accuracy for surgical gesture recognition. Results indicate that the proposed method **BML-indRNN+DCNN** outperforms the others.

IV. Conclusions and Future Works

In this paper, a BML-indRNN method is proposed for surgical gesture recognition. The proposed method is verified based on the video data of the suturing task obtained from JIGSAWS. To demonstrate the significance of the spatial feature extraction model, comparisons are made between with and without using the DCNN based spatial feature extraction model to process the image data before feeding the data into the BML-indRNN model. Results indicate that the classification accuracy can improve 4.97% when using the DCNN model with fine-tuning for spatial feature extraction. The interpretability of the feature extraction model can be verified through GradCAM for post-hoc analysis. The proposed BML-indRNN is effective, since the overall accuracy of the surgical gesture recognition for the suturing task can reach 87.13%, which outperforms several state-of-the-art approaches.

Future work will include incorporating kinematic data to further enhance the classification accuracy for surgical gesture recognition. The real-time performance of the model will be tested while more experimental validation can be conducted based on other surgical tasks, including needle passing tasks, knot-tying tasks, etc. By training effective models for surgical gesture segmentation and recognition, context-awareness for surgical operation assistance can be realized, while more reliable skill evaluation can be achieved by detailed analysis of the surgical gestures.

References

[1] C. Bergeles and G.-Z. Yang, "From passive tool holders to microsurgeons: safer, smaller, smarter surgical robots," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1565–1576, 2013.

[2] D. Zhang, Y. Guo, J. Chen, J. Liu, and G.-Z. Yang, "A handheld master controller for robot-assisted microsurgery," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 394–400.

[3] R. Miyazaki, K. Hirose, Y. Ishikawa, T. Kanno, and K. Kawashima, "A master–slave integrated surgical robot with active motion transformation using wrist axis," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 3, pp. 1215–1225, 2018.

[4] D. Zhang, J. Liu, L. Zhang, and G. Z. Yang, "Hamlyn crm: a compact master manipulator for surgical robot remote control," *International Journal of Computer Assisted Radiology & Surgery*, vol. 15, no. 3, pp. 503–514, 2020.

[5] D. Zhang, J. Liu, L. Zhang, and G.-Z. Yang, "Design and verification of a portable master manipulator based on an effective workspace analysis framework," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[6] D. Zhang, J. Liu, A. Gao, and G.-Z. Yang, "An ergonomic shared workspace analysis framework for the optimal placement of a compact master control console," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2995–3002, 2020.

[7] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, "Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions," *Computer Aided Surgery*, vol. 11, no. 5, pp. 220–230, 2006.

[8] D. Zhang, W. U. Zicong, J. Chen, A. Gao, and G. Z. Yang, "Automatic microsurgical skill assessment based on cross-domain transfer learning," *IEEE Robotics & Automation Letters*, vol. 5, no. 3, pp. 4148–4155, 2020.

[9] D. Zhang, J. Chen, W. Li, D. B. Salinas, and G.-Z. Yang, "A microsurgical robot research platform for robot-assisted microsurgery research and training," *International journal of computer assisted radiology and surgery*, vol. 15, no. 1, pp. 15–25, 2020.

[10] J. Ruurda, T. J. Van Vroonhoven, and I. Broeders, "Robot-assisted surgical systems: a new era in laparoscopic surgery." *Annals of the Royal College of Surgeons of England*, vol. 84, no. 4, p. 223, 2002.

[11] 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 & Automation Letters*, pp. 1–1, 2018.

[12] C. J. Payne, K. Vyas, D. Bautista-Salinas, D. Zhang, H. J. Marcus, and G.-Z. Yang, "Shared-control robots," *Neurosurgical Robotics*, pp. 63–79, 2021.

[13] N. Preda, F. Ferraguti, G. De Rossi, C. Secchi, R. Muradore, P. Fiorini, and M. Bonfè, "A cognitive robot control architecture for autonomous execution of surgical tasks," *Journal of Medical Robotics Research*, vol. 1, no. 04, p. 1650008, 2016.

[14] 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 banoian optimization," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 2943–2949.

[15] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel *et al.*, "Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy," *Science Robotics*, vol. 2, no. 4, p. 8638, 2017.

[16] B. van Amsterdam, H. Nakawala, E. De Momi, and D. Stoyanov, "Weakly supervised recognition of surgical gestures," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9565–9571.

[17] J. J. H. Leong, M. Nicolaou, L. Atallah, G. P. Mylonas, A. W. Darzi, and G. Z. Yang, "Hmm assessment of quality of movement trajectory in laparoscopic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2006, pp. 752–759.

[18] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse hidden markov models for surgical gesture classification and skill evaluation," in *International Conference on Information Processing in Computer-assisted Interventions*, 2012.

[19] B. Varadarajan, *Learning and inference algorithms for dynamical system models of dextrous motion*. The Johns Hopkins University, 2011.

[20] C. J. Pérez-del Pulgar, J. Smisek, I. Rivas-Blanco, A. Schiele, and V. F. Muñoz, "Using gaussian mixture models for gesture recognition during haptically guided telemanipulation," *Electronics*, vol. 8, no. 7, p. 772, 2019.

[21] G. Forestier, F. Lalys, L. Riffaud, B. Trelhu, and P. Jannin, "Assessment of surgical skills using surgical processes and dynamic time warping," 2011.

[22] E. Mavroudi, D. Bhaskara, S. Sefati, H. Ali, and R. Vidal, "End-to-end fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1558–1567.

[23] D. Ravı, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[24] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE transactions on information technology in biomedicine*, vol. 12, no. 1, pp. 20–26, 2008.

[25] F. A. Gers, D. Eck, and J. Schmidhuber, "Applying lstm to time series predictable through time-window approaches," in *Neural Nets WIRN Vietri-01*. Springer, 2002, pp. 193–200.

[26] 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.

[27] I. Funke, S. Bodenstedt, F. Oehme, F. von Bechtolsheim, J. Weitz, and S. Speidel, "Using 3d convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 467–475.

[28] Y. Qin, S. A. Pedram, S. Feyzabadi, M. Allan, A. J. McLeod, J. W. Burdick, and M. Azizian, "Temporal segmentation of surgical sub-tasks through deep learning with multiple data sources," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 371–377.

[29] F. Despinoy, D. Bouget, G. Forestier, C. Penet, N. Zemiti, P. Poignet, and P. Jannin, "Unsupervised trajectory segmentation for surgical gesture recognition in robotic training," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1280–1291, 2015.

[30] D. Liu and T. Jiang, "Deep reinforcement learning for surgical gesture segmentation and classification," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 247–255.

[31] D. Itzkovich, Y. Sharon, A. Jarc, Y. Refaely, and I. Nisky, "Using augmentation to improve the robustness to rotation of deep learning segmentation in robotic-assisted surgical data," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5068–5075.

[32] X. Gao, Y. Jin, Q. Dou, and P. A. Heng, "Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search," 2020.

[33] N. Ahmidi, L. Tao, S. Sefati, Y. Gao, C. Lea, B. B. Haro, L. Zappella, S. Khudanpur, R. Vidal, and G. D. Hager, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2025–2041, 2017.

[34] 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.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[36] J. Deng, W. Dong, R. Socher, L. J. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[37] F. Meng, K. Huang, H. Li, and Q. Wu, "Class activation map generation by representative class selection and multi-layer feature fusion," *arXiv preprint arXiv:1901.07683*, 2019.

[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2015.

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[40] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5457–5466.

[41] 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.

[42] C. Lea, A. Reiter, R. Vidal, and G. D. Hager, "Segmental spatiotemporal cnns for fine-grained action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 36–52.