Toward Gesture Recognition in Robot-Assisted Surgical Procedures

Hoangminh Huynhnguyen
Dept. of Computer Science
University of Illinois at Chicago
Chicago, USA
hhuynh20@uic.edu

Ugo A. Buy
Dept. of Computer Science
University of Illinois at Chicago
Chicago, USA
buy@uic.edu

***Abstract*—Surgical gesture segmentation and recognition are important steps toward human-robot collaboration in robot-assisted surgery. In the human-robot collaboration paradigm, the robot needs to understand the surgeon's gestures to perform its tasks correctly. Therefore, training a computer vision model to segment and classify gestures in a surgery video is a focus in this field of research. In this paper, we propose a 2-phase surgical gesture recognition method and we evaluate empirically the method on JIGSAWS's suturing video dataset. Our method consists of a 3D convolutional neural network to detect the transition between 2 consecutive surgemes and a convolutional long short-term memory model for surgeme classification. To the best of our knowledge, ours is the first study aimed at detecting action transition in a multi-action video and to classify surgemes using an entire video portion rather than classifying individual frames. We also share our source code at https://github.com/jemiar/surgery-gesture-recog**

***Index Terms*—Gesture recognition, video segmentation, robot-assisted surgery, deep learning, 3D CNN, Convolutional LSTM.**

I. Introduction

Robot-assisted surgery has improved the quality of surgery and enabled radical innovations such as tele-surgery. Compared to traditional open surgery, robot-assisted surgery results in smaller incisions, reduced bleeding and infection risk, shorter hospital stays and speedier recoveries [1], [2]. In addition, robot-assisted surgery improves surgeons' dexterity by providing 7 degrees of freedom on 4 robot arms, reduces fatigue, and improves precision by eliminating the tremor inevitably associated with human exertion [1].

However, robot-assisted surgery requires specialized expertise from surgeons. Between 150 and 250 training procedures are typically needed in order for a surgeon to reach proficiency in manipulating robotic controls [2]. Due to the lack of haptic feedback on robot motions as well as the demands of hand-eye coordination through video, performing common surgical tasks such as suturing can be tedious and time-consuming [1].

To mitigate these issues, various research efforts seek to automate simple and repetitive surgical tasks, including suturing and dissecting. Surgical robots can be taught to learn from expert demonstration. For instance, Reiley et al. used Gaussian Mixture Models to capture the trajectory of surgical effectors, and then generated new trajectories for the robot using Gaussian Mixture Regression [3]. Schulman implemented a non-rigid registration to transform a trajectory from a demonstration scene to a test scene [4]. Finite State Machines can help teaching a robot about the transition graph between surgemes [5]. Recently reinforcement learning has been applied to simple tasks, including tissue manipulation [6], [7], and needle insertion [8].

While current robots are entirely passive devices under the control of a human physician, achieving full automation is quite challenging because (1) surgical robots modify their working environment by moving organs during procedures and (2) predicting the interactions between human organs and surgical effectors is rather difficult. In the current state of the art, partial automation is a more realistic objective whereby robots and human collaboratively perform surgical tasks. In this line of research, Watanabe devised a suturing method in which a surgeon instructs the robot on the high-level features of a suturing task (e.g., the locations where suturing starts and ends) but the low-level sub-tasks (e.g., needle pulling and hand-over between arms) are performed autonomously by the robot [9].

Critically for an effective human-robot collaboration in surgery, the robot must be able to understand the surgeon's gestures by using kinematic data measured on the hardware controls, or video data, or both. JIGSAWS [10] is one of the best-known and widely used datasets for surgical gesture recognition as it provides a comprehensive labeled kinematic and video dataset for robot-assisted surgery. In their literature review on gesture recognition for robotic surgery [11], van Amsterdam et al. summarized a many machine learning approaches in this area: Hidden Markov Models (HMM), Dynamic Linear System (DLS), Conditional Random Fields (CRF), reinforcement learning, Convolutional Neural Nets (CNNs), Recurrent Neural Networks (RNNs) as well as unsupervised and semi-supervised learning techniques.

The deep learning models in [11] mostly used the encoder-decoder structure and applied it to individual frames. In this paper, we define a 2-stage deep learning approach to recognizing surgical gestures from video data in the JIGSAWS data set. Stage 1 involves determining whether a block of 10 consecutive frames sampled at a frequency of 10 Hz is a transition block between 2 surgemes or a block belonging

978-1-7281-9690-9/20/$31.00 © 2020 IEEE

to the same surgeme using a 3D CNN. In Stage 2, blocks of frames between transition are classified into 10 kinds of surgemes using a hybrid model combining a CNN and a Long Short-Term Memory (LSTM) net. We hypothesize that by classifying the whole block of frames in Stage 2, our model can preserve both spatial and temporal information from the video data, and can properly learn the characteristic of each surgeme for more precise gesture classification.

This paper is organized as follows. We will introduce related work in gesture classification and recognition in Section 2. We will then present our network architectures, experimental setups, results, and discussion in Sections 3 through 6.

II. Related Work

A. Gesture Classification

Gesture classification solves the problem of classifying trimmed video or kinematic data, or both into corresponding classes of surgeme. Haro et al. used LDS, Bag-of-Features (BoF) and their combination, along with Support Vector Machines (SVM) to classify surgemes in the JIGSAWS data set [12]. They then improved their result by using both video and kinematic data in [13]. Thanks to well-defined classes of surgemes, in [14] Fard applied k-Nearest Neighbor (kNN) techniques to kinematic data transformed with Dynamic Time Warping (DTW). In the era of Deep Learning, Luongo et al. used Convolutional LSTMs on real surgery video data to classify needle-driving vs. non-needle-driving gestures [15]. With the same Convolutional LSTM architecture, Sarikaya [16] applied multimodal learning by combining video data and optical flow to classify gestures and tasks on the JIGSAWS data set.

B. Gesture Recognition

Unlike the gesture classification task, the gesture recognition task must read the entire recording (video, kinematic or both) of a surgical procedure, and then try to recognize where each surgeme starts and ends in that procedure. Initial works in this area used graphical models, such as HMM and CRF to learn the transition graph between surgemes, and then use this graph to classify each frame in the surgery recording [17], [18]. The development of deep learning models has helped accelerate works in gesture recognition. Multiple deep learning architectures based on CNNs, RNNs, their variants, and hybrid combinations have been proposed and validated on the JIGSAWS dataset. Studies that used video data as input for their model usually relied on CNN architecture, such as 3D CNN [19], Temporal Convolutional Networks [20] and Spatiotemporal CNNs [21]. The RNN architecture was applied to kinematic data [22]. CNNs and RNNs have been combined in hybrid models, such as the TricorNet [23]. Despite having different architectures, these deep learning models shared an encoder-decoder design, and they all aimed to classify each individual frame in the surgery recording. Reinforcement learning has also been applied in surgical gesture recognition. Liu trained a reinforcement learning model that could learn a policy to classify frames of a surgical procedure [24].

Our work can be categorized as a deep learning solution to gesture recognition. In contrast with existing work, our solution does not use the encoder-decoder design. Instead we seek to analyze a whole video segment covering a surgical task, such as an entire suturing action. To facilitate this process we use a 3D CNN to detect the transition between surgemes, and then use a Convolutional LSTM to classify each surgeme. We hypothesize that by classifying each surgeme as an entity, as opposite to classifying each frame, our network can preserve both spatial and temporal information, and learn the characteristic of individual surgemes.

III. Network Architectures

We use two distinct architectures for our two goals. The first architecture seeks to distinguish video segments containing a transition between two consecutive surgemes from segments that do not contain such a transition. The second architecture seeks to classify the video segments between transitions (i.e., video segments that do not contain a surgemes) into 10 different classes of subtasks contained in a suturing task. The two network architectures are discussed next.

A. Gesture Transition Classification Model

For the gesture transition classification model, we first apply 3 layers of a 3D CNN with 8, 16, 32 filters respectively to the input, as depicted in Figure 1. In our code, each 3D convolutional layer is accompanied by a max pooling layer. After the 3 convolutional layers, we use an average pooling layer to flatten the activation matrix. We then add a dense layer with 512 nodes before the output layer. Here the input consists of 10 frames with 5 frames preceding the transition and 5 frames following the transition from one surgeme to the next. Frames are sampled at 10 Hz.

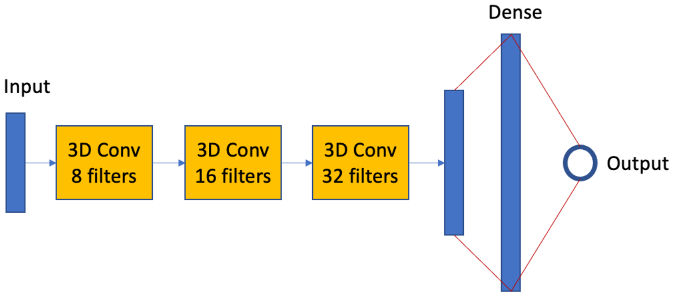

Fig. 1. Gesture transition classification model.

B. Gesture Classification Model

Since each gesture has different length, we use a Convolutional LSTM net for the gesture classification task as depicted in Figure 2. For each frame in the input, we apply a 2D convolutional layer twice. As with gesture transition classification above, we also have a max pooling layer after each convolutional layer. After the 2 layers of 2D CNN, we flatten the outputs before routing them to the LSTM layer. The output of the last LSTM block is input into a dense layer with 32 nodes. The output is a dense layer with 10 nodes, representing the 10 surgemes defined in the JIGSAWS suturing data set. We use the Keras library and their code examples to design our 3D CNN and Convolutional LSTM models [25].

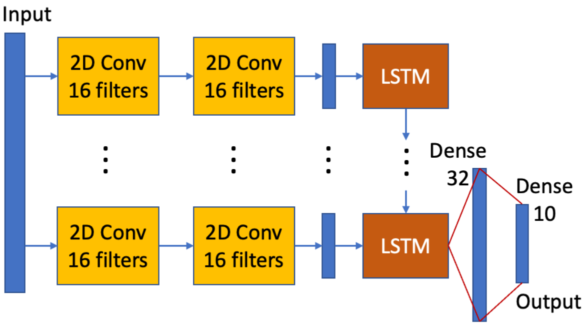

Fig. 2. Gesture classification model.

IV. Experiments

We used the suturing portion of the JIGSAWS dataset [10] for our study. The original video data were recorded at 30Hz with 480x640 pixels per frame. The resulting dataset contains 39 videos total, which were recorded by 8 surgeons with different levels of proficiency performing the suturing task in 5 trials each. We downsampled the video to 10Hz, and resized frames to 240x320 pixels to accelerate the training process.

For gesture transition classification, we collected blocks of 10 consecutive frames. A transition block consists of 5 frames belonging to the first surgeme followed by 5 additional frames belonging to the next surgeme. A normal block has all its 10 frames belonging to the same surgeme. Evidently, our dataset contains many more normal blocks (3940 blocks) than transition blocks (754 blocks). To address this issue, we upsampled the transition blocks and downsampled the normal blocks, resulting in a training dataset of 600 normal blocks and around 500 transition blocks. We ran the validation process on 200 normal blocks and 150 transition blocks.

In the gesture classification task, we used the transcript contained in JIGSAWS to clip the video data for each surgeme, and then fed each surgeme to our classification model. There are 10 surgemes in the suturing data set. The gestures and their occurrences are listed in Table I. For gestures that occurred less frequently, such as G1, G5, G8, G9 and G10, we upsampled them to around 100 in order to obtain a balanced training dataset. We kept the original occurrences of gestures in the validation data set.

We used Leave-One-Supertrial-Out (LOSO) for 5-fold cross-validation in both tasks, as recommended elsewhere [10]. In gesture transition classification, we used mini-batch gradient descent with size 16. In gesture classification, we used Stochastic Gradient Descent (SGD) because the input samples were of different length. As we dealt with a large video dataset, we follow well-established procedures [26] to load the video data into our models.

TABLE I
Gestures In the Suturing Data set

Gesture	Gesture Description	Occurrences
G1	Reaching for needle with right hand	29
G2	Positioning needle	166
G3	Pushing needle through tissue	164
G4	Transferring needle from left to right	119
G5	Moving to center with needle in grip	37
G6	Pulling suture with left hand	163
G8	Orienting needle	47
G9	Using right hand to help tighten suture	24
G10	Loosening more suture	4
G11	Dropping suture and moving to end points	39

V. Results

After running LOSO cross-validation for the gesture transition classification task, we achieved a result of around 70% for accuracy, precision and recall, as shown in Table II.

TABLE II
Gesture Transition Classification Result

Trial	Accuracy	Precision	Recall	F1 Score
Trial 1	71.74%	71.01%	68.57%	0.70
Trial 2	69.49%	66.67%	66.13%	0.66
Trial 3	75.89%	71.63%	71.13%	0.71
Trial 4	71.13%	63.69%	74.83%	0.69
Trial 5	77.08%	71.72%	74.29%	0.73
Average	73.07%	68.94%	70.99%	0.70

Similarly, after running the cross-validation for the gesture classification, we achieved an average accuracy of 76.3%, as shown in Table III. The precision, recall and F1 score of 10 classes of gestures is presented in Table IV. We were not able to calculate precision, recall and F1 score for G10 due to low occurrence frequency.

TABLE III
Gesture Classification Result

Trial	Accuracy
Trial 1	72.63%
Trial 2	78.95%
Trial 3	75.32%
Trial 4	73.42%
Trial 5	81.17%
Average	76.30%

VI. Discussion

In this study, we used a 3D CNN and a Convolutional LSTM to detect transitions between surgical gestures, and to classify the gestures. We used these models to learn the movement of each gesture via video data exclusively. We find our current results quite encouraging in consideration of the following challenges. First, because the JIGSAWS dataset included surgeons with different levels of proficiency, their gesture movements were likely different from each other which adversely affected our overall accuracy figures. If we

TABLE IV
PRECISION, RECALL, F1 SCORE OF 10 SURGICAL GESTURES

Gesture	Precision	Recall	F1 Score
G1	79.17%	73.22%	0.76
G2	72.01%	83.38%	0.77
G3	86.38%	75.02%	0.80
G4	60.76%	67.69%	0.64
G5	33.33%	27.78%	0.30
G6	86.68%	92.27%	0.89
G8	49.41%	43.28%	0.45
G9	100%	75%	0.88
G10	N/A	N/A	N/A
G11	93.76%	93.76%	0.94

had a larger dataset from the same surgeon, or from surgeons with high level of proficiency, we would have been able to achieve a greater accuracy results. We can also pre-classify surgeons, using [27] before running our model to get a better result.

Second, the size JIGSAWS dataset is probably inadequate from the viewpoint of statistical learning. Some of the 10 surgemes that we classified occurred less frequently than others in the suturing procedure. The low frequency of these gestures might have reduced the performance of our models. When we look at Table I and Table IV, gestures with low frequency, such as G5, G8 and G10 had a much lower precision, recall and F1 score than other gestures. In this light, we find our 76% accuracy benchmark to be quite satisfactory, considering that this is a first effort in detecting surgical gesture transitions.

VII. FUTURE WORKS

In this study, we define a 2-stage approach to surgical gesture segmentation and recognition: Stage 1 detects transition gestures and Stage 2 classifies video clips into correct gesture classes. We have achieved an accuracy of over 70% for both tasks. In the next step, we will seek to improve the performance of both models by collecting more data and by applying regularization. We will also combine the 2 stages together to see how they work and compare their results with other studies in surgical gesture recognition.

ACKNOWLEDGMENT

This study is supported in part by a 2020 Discovery Partners Institute Science Team Seed Grant.

REFERENCES

[1] Lanfranco, Anthony R., et al. ”Robotic Surgery: A Current Perspective.” Annals of Surgery 239.1 (2004): 14.

[2] Barbash, Gabriel I. ”New Technology and Health Care Costs–The Case of Robot-Assisted Surgery.” The New England Journal of Medicine 363.8 (2010): 701.

[3] Reiley, Carol E., Erion Plaku, and Gregory D. Hager. ”Motion Generation of Robotic Surgical Tasks: Learning from Expert Demonstrations.” 2010 Annual international conference of the IEEE engineering in medicine and biology. IEEE, 2010.

[4] Schulman, John, et al. ”A Case Study of Trajectory Transfer through Non-Rigid Registration for a Simplified Suturing Scenario.” 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2013.

[5] Sen, Siddarth, et al. ”Automating Multi-Throw Multilateral Surgical Suturing with a Mechanical Needle Guide and Sequential Convex Optimization.” 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2016.

[6] Thananjeyan, Brijen, et al. ”Multilateral Surgical Pattern Cutting in 2D Orthotropic Gauze with Deep Reinforcement Learning Policies for Tensioning.” 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017.

[7] Shin, Changyeob, et al. ”Autonomous Tissue Manipulation via Surgical Robot Using Learning Based Model Predictive Control.” 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019.

[8] Keller, Brenton, et al. ”Optical Coherence Tomography-Guided Robotic Ophthalmic Microsurgery via Reinforcement Learning from Demonstration.” IEEE Transactions on Robotics 36.4 (2020): 1207-1218.

[9] Watanabe, Kengo, et al. ”Single-Master Dual-Slave Surgical Robot with Automated Relay of Suture Needle.” IEEE Transactions on Industrial Electronics 65.8 (2017): 6343-6351.

[10] Gao, Yixin, et al. ”Jhu-isi Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling.” MICCAI Workshop: M2CAI. Vol. 3. 2014.

[11] van Amsterdam, Beatrice, Matthew Clarkson, and Danail Stoyanov. ”Gesture Recognition in Robotic Surgery: A Review.” IEEE Transactions on Biomedical Engineering (2021).

[12] Haro, Benjamín Béjar, Luca Zappella, and René Vidal. ”Surgical Gesture Classification from Video Data.” International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, Heidelberg, 2012.

[13] Zappella, Luca, et al. ”Surgical Gesture Classification from Video and Kinematic Data.” Medical Image Analysis 17.7 (2013): 732-745.

[14] Fard, Mahtab J., et al. ”Distance-based Time Series Classification Approach for Task Recognition with Application in Surgical Robot Autonomy.” The International Journal of Medical Robotics and Computer Assisted Surgery 13.3 (2017): e1766.

[15] Luongo, Francisco, et al. ”Deep Learning-based Computer Vision to Recognize and Classify Suturing Gestures in Robot-Assisted Surgery.” Surgery (2020).

[16] Sarikaya, Duygu, Khurshid A. Guru, and Jason J. Corso. ”Joint Surgical Gesture and Task Classification with Multi-Task and Multimodal Learning.” arXiv preprint arXiv:1805.00721 (2018).

[17] Tao, Lingling, et al. ”Surgical Gesture Segmentation and Recognition.” International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, Heidelberg, 2013.

[18] Lea, Colin, Gregory D. Hager, and Rene Vidal. ”An Improved Model for Segmentation and Recognition of Fine-Grained Activities with Application to Surgical Training Tasks.” 2015 IEEE Winter Conference on Applications of Computer Vision. IEEE, 2015.

[19] Funke, Isabel, et al. ”Using 3D Convolutional Neural Networks to Learn Spatiotemporal Features for Automatic Surgical Gesture Recognition in Video.” International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019.

[20] Lea, Colin, et al. ”Temporal Convolutional Networks: A Unified Approach to Action Segmentation.” European Conference on Computer Vision. Springer, Cham, 2016.

[21] Lea, Colin, et al. ”Segmental Spatiotemporal CNNs for Fine-Grained Action Segmentation.” European Conference on Computer Vision. Springer, Cham, 2016.

[22] Gurcan, Ilker, and Hien Van Nguyen. ”Surgical Activities Recognition Using Multi-scale Recurrent Networks.” ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[23] Ding, Li, and Chenliang Xu. ”Tricornet: A Hybrid Temporal Convolutional and Recurrent Network for Video Action Segmentation.” arXiv preprint arXiv:1705.07818 (2017).

[24] Liu, Daochang, and Tingting Jiang. ”Deep Reinforcement Learning for Surgical Gesture Segmentation and Classification.” International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018.

[25] https://keras.io/examples/vision/

[26] https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly

[27] Getty, Neil, et al. ”Recurrent and Spiking Modeling of Sparse Surgical Kinematics.” International Conference on Neuromorphic Systems 2020. 2020.