Gesture Classification in Robotic Surgery using Recurrent Neural Networks with Kinematic Information

E.B. Mazomenos[1], D. Watson[2], R. Kotorov[2] and D. Stoyanov[1]

[1] *Wellcome/EPSRC Centre for Interventional and Surgical Sciences, Department of Computer Science, University College London, London, U.K.*

[2] *Trendalyze Inc., London, U.K.*

INTRODUCTION

The integration of robotics in minimally-invasive surgery has witnessed remarkable increase over the previous decade. Breakthrough innovations in robotic technology, imaging and sensing facilitated the design of novel surgical systems for a number of different operations (laparoscopy, endovascular surgery). Prime example is the da Vinci Surgical System (dVSS; Intuitive Surgical Inc., Sunnyvale, CA, USA) used nowadays in many laparoscopic resection procedures (prostatectomy, cholecystectomy, nephrectomy) while it is constantly expanding to other surgical domains.

The use of robotic technology offers significant operational advantages like increased maneuverability, reduction of tremor and more precise tool positioning thus minimising intra-operative risk and trauma ultimately leading to a reduction in recovery times [3]. The continuous development of image-guided robotic surgery creates a need for new surgeons to go through analogous training for this type of surgery in order to master the necessary dexterous and technical skills. The currently practiced method of surgical training is heavily-based on expert supervision, with faculty surgeons reviewing and evaluating performance through manually assessing global rating scales and task specific checklists. The scoring procedure requires significant amount of time and it is also subjective and prone to interobserver variability. Subsequently, it has been advocated that novel objective methods, focusing on competency metrics should be developed for evaluating surgical trainees. Typically, robotic surgical systems, like the dVSS, have the ability to record both video and tool kinematic parameters (joints pose). This offers the possibility for analysing surgical procedures and developing objective performance methods based on the manipulation pattern of surgical tools. Procedures can be broken down to sequential surgical tasks which can be further partitioned to autonomous activities termed as "gestures". It has been reported that the ability to recognize surgical gestures can be further exploited for performance assessment [1].

In this work we introduce the application of Recurrent Neural Networks (RNNs) on surgical kinematic data, for the classification of gestures in three fundamental surgical tasks (suturing, needle passing knot tying). The developed RNN-based classifier achieves close to 60% average classification accuracy for all three tasks when trained and tested with dVSS kinematic data from the same operator. Our preliminary work indicates that this type of artificial neural networks can be the building blocks in gesture classification systems which can form the basis for further developing automated skill assessment methods in robotic surgery.

MATERIALS AND METHODS

The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) is a publicly available surgical dataset comprising of video and kinematic data from the execution of three basic surgical tasks (suturing, knot tying and needle passing) with the dVSS on bench-top models by eight surgeons (subjects) of varying level of expertise [1, 2]. All subjects performed each task five times. The stereo video output of the dVSS endoscopic camera module was captured at 30fps in 640x480 resolution. The kinematic data contain 3D position, orientation, velocity and gripper angle values from both the master and slave, left and right manipulators totaling 76 motion-related parameters. The two datastreams are synchronised with the same sampling rate.

Figure. 1. The three surgical tasks performed in JIGSAWS: from left to right – Suturing; Needle Passing; Knot Tying.

A vocabulary of subtasks (gestures) is also formulated for representing each task in JIGSAWS. A surgical gesture is considered as a single action that completes a clearly identifiable step. Gestures are completed sequencially, and their entire sequence comprises the overall task. Fifteen different gestures are defined in JIGSAWS and used to manually annotate the dataset in such a way that each temporal datapoint (video and kinematics) is assigned a single gesture. The list of gestures in JIGSAWS is:

(G1) reaching for the needle with right hand;
(G2) positioning the tip of the needle;
(G3) pushing needle through the tissue;
(G4) transferring needle from left to right;
(G5) moving to center of workspace with needle in grip;
(G6) pulling suture with left hand;

(G7) pulling suture with right hand;
(G8) orienting needle;
(G9) using right hand to help tighten suture;
(G10) loosening more suture;
(G11) dropping suture and moving to end points;
(G12) reaching for needle with left hand;
(G13) making C loop around right hand;
(G14) reaching for suture with right hand;
(G15) pulling suture with both hands.

RNNs is a class of neural networks that their structure includes directed connections along a sequence, like a graph, allowing information to persist. Each node in a an RNN has a time-varying activation value and each connection between nodes carries a modifiable weight. A standard building block of RNNs is a Long-Short Term Memory (LSTM) unit. This comprises of the input and output stages and the internal cell as illustrated in Figure 2.

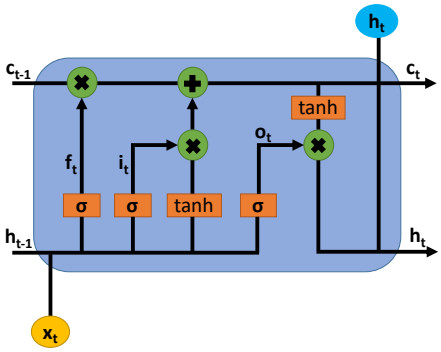

Figure. 2. The fundamental LSTM unit.

For our multi-label classification problem we design an RNN using an LSTM block with an internal dimension of 128 followed by a fully-connected (FC) layer with softmax activation that concatenates the LSTM output and produces the classification result. The dimension of the FC layer is set to the number of gestures that we intend to classify. To minimise overfitting we employ dropout in the LSTM block with a value of 0.3. Our investigation takes place with the kinematic information of each subject separately for each task. This resulted to the dense layer having variable dimension since each subject may perform the task using a different number of gestures and in situations gestures not present in the nominal sequence as defined in JIGSAWS.

Our intention was to initially evaluate the ability of the RNN to classify the surgical gestures of each individual subject, in the three tasks, only using data from that particular subject. We therefore collated the kinematics from all executions that a subject performed on each task and used the 80%-20% rule to separate the data into a training and testing dataset. Through experimentation we identified that the inclusion of the gripper angle value diminishes performance, hence we chose to disregard this parameter from both the master and slave and perform our investigation with the remaining 72 kinematic variables. The gesture annotations were used as the ground truth labels. For training the RNN the cross-entropy was set as the loss function and gradient descent optimization with adaptive moment estimation was performed to obtain the weights of the LSTM connections. The RNN network was trained for 15 epochs with a batch size of 64.

RESULTS

The performance of the RNN was evaluated using as accuracy metric the percentage of correct gesture classification over the total number of gesture annotations. Table 2 lists per subject and average accuracy results for the three tasks.

Table 1. Classification accuracy results (subject 6 had no annotations for the Needle Passing task)

Subject	Suturing	Needle Passing	Knot Tying
1	63.22%	72.20%	57.28%
2	63.10%	70.89%	82.70%
3	75.70%	36.86%	69.23%
4	64.50%	56.80%	56.15%
5	64.49%	79.49%	24.03%
6	59.80%	-	76.18%
7	67.66%	47%	52.05%
8	73.69%	76.76%	58.73%
Average	66.52%	62.85%	59.54%

CONCLUSION AND DISCUSSION

In this paper we have demonstrated that RNNs have considerable potential as building blocks in robotic surgical gesture classification systems. The ability to partition surgical tasks into simple gestures can be exploited in the development of objective performance assessment methods. Results show that close to 60% average classification accuracy can be achieved with a simple RNN-based gesture classifier. Future efforts will focus on boosting classification performance by developing hybrid, combining Convolution Neural Networks and RNNs, learning-based systems that combine both kinematic and video information.

REFERENCES

[1] N. Ahmidi et al., "A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery", in IEEE Trans Biomed Eng, vol. 64 (9), pp. 2025-2041, Sep. 2017.

[2] Yixin Gao et al., "The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling", In Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop, vol 3 2014.

[3] C. Bergeles and G. Z. Yang, "From Passive Tool Holders to Microsurgeons: Safer, Smaller, Smarter Surgical Robots", in IEEE Trans Biomed Eng, vol. 61(5), pp. 1565-1576, May 2014.